IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of: Grasso, L. *et al.*	**Confirmation No.:** 9935
Application No.: 10/624,631	**Group Art Unit:** 1633
Filing Date: July 21, 2003	**Examiner:** Kevin Hill, Ph.D.

Title: Methods For Generating Enhanced Antibody-Producing Cell Lines With Improved Growth Characteristics

Mail Stop AF
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

DECLARATION OF J. BRADFORD KLINE, PH.D. PURSUANT TO 37 CFR § 1.132

I, J. Bradford Kline, being duly warned that willful false statements and the like are punishable by fine or imprisonment or both, under 18 U.S.C. § 1001, and may jeopardize the validity of the patent application or any patent issuing thereon, state and declare as follows:

1. All statements herein made of my knowledge are true and statements made on information or belief are believed to be true. The Exhibits attached hereto are incorporated herein by reference.

2. I received my Ph.D. degree in Microbiology & Immunology in 1996 at the University of Miami School of Medicine. I received my B.A. in Biology from the University of Virginia in 1989.

3. I have been employed by Morphotek for six years. I am an Assistant Director of Research and Development. I am currently head of the Genomics/Ab Core Group and responsible for the recombinant expression of our therapeutic lead candidates. I am an inventor of the subject matter pending in U.S. Pat. Appl. 10/624,631. A copy of my *curriculum vitae* is attached hereto as Exhibit A.

4. I have read and am familiar with the contents of the above-referenced patent application, and the Office Action dated April 4, 2007. I understand that the nature of the rejection at issue in the pending application is that the Office has asserted that the pending claims fail to meet the written description and enablement requirements. As part of that rejection, the Office has argued that the application's disclosure does not sufficiently enable the claims to alpha-1-antitrypsin (*AAT*) and endothelial monocyte-activating polypeptide I (*EMAP*) knockout cell lines. The Office's argument is based primarily on two principle considerations, the unpredictability of the field of gene knockout, and a failure to adequately disclose gene knockout starting materials. The purpose of this declaration is to address these issues.

5. As of this application's filing, gene knockout technology was wide-spread, documented, and predictable. Given the information in this application's disclosure, and in light of the state of the art, an ordinary scientist in the field could have produced alpha-1-antitrypsin (*AAT*) and endothelial monocyte-activating polypeptide I (*EMAP*) knockout cells without undue or excess experimentation. The extent of experimentation required to practice the invention is further limited because several of the antisense methods that we employed share common techniques and reagents with the established methods for gene knockout. One of skill in the art would also expect the *AAT* and *EMAP* antisense phenotypes described in the application to be the same as the *AAT* and *EMAP* knockout phenotypes. Moreover, at the time of this application's filing, there was an expectation of success for knockout of the *AAT* and *EMAP* genes based on the experiments described in the Examples section of the specification.

Scientific Literature Addressing Gene Knockout Technology

6. The first gene knockout cells were produced during the 1980s. By 2003, thousands of published reports documented gene knockout technology and methods. In addition, as of 2003, there were several standard reference works setting forth the general principles of recombinant DNA technology, including: Ausubel *et al.* CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, John Wiley & Sons, New York (1998); Sambrook *et al.* MOLECULAR CLONING: A LABORATORY MANUAL, 2D ED., Cold Spring Harbor Laboratory Press, Plainview, New York (1989); Kaufman *et al.*, Eds., HANDBOOK OF MOLECULAR AND CELLULAR METHODS IN BIOLOGY AND MEDICINE, CRC Press, Boca Raton (1995); McPherson, Ed., DIRECTED MUTAGENESIS: A PRACTICAL

APPROACH, IRL Press, Oxford (1991). The technology set forth in these references was, and still is, readily practiced in gene knockout technology.

7. By 2003, multiple books had been published specifically addressing the technique of gene knockout, many including specific experimental details for producing gene knockout cells and organisms. Tymms *et al.*, Eds., GENE KNOCKOUT PROTOCOLS, Humana Press (2001), is one such book, and provides highly detailed protocols for the production of gene knockouts, including reagent concentrations, reaction conditions, times, and temperatures, and screening methods. Given the advancement of gene knockout technology in the scientific community and the extent of literature addressing recombinant technology and gene knockout protocols, it would have been straightforward for an ordinary scientist in the field to produce *AAT* and *EMAP* knockout cells from this application's disclosure.

8. I disagree with the Examiner's assertion that the field of gene knockout was unpredictable as of 2003, particularly as it relates to the present invention. Gene knockout was predictable as of 2003 because the methods had been in use for nearly two decades, thousands of articles had been published, and there were detailed protocols for performing the methods. The high degree of success scientists have demonstrated in the field for widely divergent genes, cells, and organisms, supports this assertion. Furthermore, the likelihood of success in knocking out a given gene would be expected to be greater where, as demonstrated in this application, antisense

experiments involving the same genes had been successfully performed before attempting to produce knockouts of the same genes.

Typical Gene Knockout Methods

9. As of 2003, there were numerous strategies for producing gene knockout (null allele) cells and organisms. At that time, scientists routinely employed homologous recombination for either deletion of or other mutation of a target gene. Gene targeting is the process by which a defined modification is introduced at a specific genomic location by homologous recombination. Gene targeting is achieved by transfecting a cell with specific gene-targeting vectors. In homologous recombination, a double crossover event irreversibly replaces a target gene sequence with a targeting vector.

10. The general steps involved in homologous recombination for gene knockout include: (A) ascertaining the gene sequence to be knocked out, (B) developing a gene targeting strategy, (C) producing a targeting vector that is designed to recombine with and mutate a specific chromosomal locus, (D) introducing the vector into the cell for recombination, and (E) screening the cell for effective knockout.

(A) Ascertaining the Genetic Sequence to be Knocked Out

11. In order to perform gene knockout, the sequence of the gene of interest must be known. Furthermore, the upstream and downstream sequences must also be known for production of a

homologous vector. As of 2003, a vast number of genetic libraries providing the sequences of genes of interest were available to scientists, thus simplifying this step and allowing the application of gene knockout technology to multiple organisms. From these libraries, it was straightforward in 2003 to obtain the sequence of a gene of interest and the gene's upstream and downstream flanking sequences. In fact, the *AAT* and *EMAP* genes from species such as humans, chimps, mice and rats were known and publicly available as of 2003.

(B) Developing a Gene Targeting Strategy

12. Successful gene knockout requires an effective targeting strategy. Several sources are available to assist even novices in producing strategies with high probabilities of success. (see, *e.g.*, Siew-Sim *et al.* (2001) Mol. Biotechnol. 19: 297-304; attached as Exhibit B). These methods were so well established in 2003 that college students routinely performed them in their laboratory courses. To ensure successful knockout, each target vector should contain: regions homologous to the target and the sequences flanking it, a positive selection marker, a negative selection marker, a site of targeting vector linearization, and restriction enzyme sites for screening.

13. The targeting vector should be homologous to the regions flanking the target gene. The homologous regions should be between 5 and 8 kb with roughly one half in each of the 5' and 3' vector arms. These homologous regions will flank the gene to be deleted.

14. Three major processes occur when a population of cells is transfected with a targeting vector: (1) the vector does not incorporate, (2) the vector randomly incorporates, or (3) the vector incorporates through homologous recombination at the gene target. The use of positive and negative selection vectors allows one to screen against the first two cases, leaving only cells that have undergone homologous recombination.

15. Positive selection markers incorporate along with the targeting vector and impart immunity to the cells against certain reagents. The simplest and most commonly used positive selection marker is neomycin phosphotransferase (*neo*), which imparts immunity against neomycin and kanamycin. An alternate positive selection marker is hygromycin B phosphotransferase gene (*hyg*), which provides immunity against hygromycin. Although it is rarely used, the Zeocin resistance gene (*zeo*) may also be used as a positive selection marker. Only cells incorporating the vector will survive when exposed to the reagent to which immunity has been provided by the positive selection marker. Cells randomly incorporating the vector will not be eliminated through the use of a positive selection marker.

16. Negative selection markers are used to screen against cells randomly incorporating the targeting vector. Where homologous recombination has occurred, the negative selection marker will most often be excluded from the incorporating vector. Because the negative selection marker lies outside the region of homology, it does not incorporate into the targeted gene during homologous recombination. When the targeting vector randomly incorporates, however, the

negative selection marker will frequently incorporate along with the other portions of the vector. Therefore, the negative selection marker is much more likely to be found in a cell's genome where the vector is randomly incorporated. Herpes simplex virus thymidine kinase (HSVTK)is frequently used as negative selection marker. When HSVTK is used, cells not incorporating the targeting vector through homologous recombination will not survive when treated with gancyclovir or FIAU (1-2-deoxy-2-fluroro-1-β-D-arabinosofuranosyl-5-iodouracil). An alternate negative selection marker is the diphtheria toxin A-chain (DT-A) gene. Because DT-A expression is toxic to cells, negative selection occurs without additional reagents when the DT-A negative selection marker is used.

17. Bacterial plasmids are frequently employed in homologous recombination. When plasmids are used, the circular DNA must be cleaved to produce linear DNA prior to entering the target cell. A targeting vector linearization site is incorporated into the vector, which may be cleaved by restriction enzymes to yield a linear product. In general, the restriction site is located outside of the homologous region and usually where the 3' arm attaches to the plasmid backbone. Other restriction sites are incorporated into the targeting vector to enable genome screening by Southern blot analysis following vector transfection.

18. Proper planning and incorporation of the above vector components dramatically improve the likelihood of successful gene knockout. Producing an effective targeting strategy was, however, considered routine as of 2003.

(C) Producing the Targeting Vector

19. Bacterial plasmids are frequently used to produce targeting vectors for homologous recombination because of the relative ease with which an artificially introduced mutant DNA segment will replace bacterial DNA. Production of targeting-vector clones in bacterial plasmids following transfection is also straightforward. After their production, the targeting vector plasmids are linearized and then introduced into the target cell. Vectors from yeast, insect, and mammalian cells, as well as viral vectors could also be used.

(D) Introducing the Targeting Vector into the Target Cell

20. There are several straightforward methods for introducing targeting vectors into cells. Vectors may be microinjected into mammalian cells with a glass micropipette. Electroporation may be used, whereby a brief shock causes the cell membrane to become temporarily permeable to targeting vectors. Viruses may be engineered to carry the vector into the cell. Vectors may also be introduced into plant cells through particle bombardment, whereby DNA samples are painted onto tiny gold beads and "shot" through the cell wall. All such methods were routinely practiced at the time this invention was made.

(E) Screening the Cell for Effective Knockout

21. Positive and negative selection markers allow cells to be screened on the basis of proper targeting vector incorporation through homologous recombination. Once the selectable markers

are placed in the targeting vector as described above, the cells may be treated with specific reagents to screen for proper incorporation of the targeting vector. Where *neo* is used as a positive selection marker, neomycin and kanamycin resistance would be found in cells incorporating the targeting vector. Where herpes simplex virus thymidine kinase is used as negative selection marker, cells not incorporating the targeting vector through homologous recombination will not survive when treated with gancyclovir. Moreover, knockout of more than one genetic loci in a single system was straightforward as of 2003. Multiple gene knockout is achieved where one selectable marker, such as *neo*, is employed during knockout of the first gene, and a different selectable marker, such as *hyg*, is employed for knockout of the second gene. The insertion of restriction sites into the vector permits further screening of treated cells for vector incorporation by Southern blot analysis.

Gene Knockout Method Variants

22. Gene knockout may be accomplished through several variants of the methods described above. Some of these variants include the use of the Cre-*loxP* system, the use of conditional gene knockouts, and employment of various mutation types.

Cre-Mediated Excision of Selectable Marker Elements

23. Because positive selection markers such as *neo* may alter gene expression at the targeted loci, it is often desirable to remove that selectable marker following incorporation of the targeting vector. The Cre-*loxP* system enables removal of that selectable marker. Cre

(cyclization recombination) recombinase recognizes the *loxP* site. Where two *loxP* sites flank a DNA segment, Cre will excise any intervening DNA. With this method, the *neo* marker (or other selectable marker) is flanked by two *loxP* sequences (floxed), and Cre recombinase subsequently removes the *neo* marker. The floxed-*neo* marker may be removed through transient Cre expression or by inserting a *Cre* expression vector. The Cre-*loxP* system was well established in the art as of 2003.

Conditional Knockout

24. In certain situations, it may be preferable to make gene knockout conditional. Knockout may be made conditional on factors such as temperature, or may be performed using the Cre-*loxP*-induced conditional knockout system. Conditional knockout using the Cre-*loxP* system occurs in the same manner as described in the preceding section, except the gene of interest (or some functionally important portion of it) is flanked by *loxP* and excised upon expression of Cre recombinase.

Mutation Type Variations

25. As of 2003, gene knockout could be achieved using various gene mutation types. Gene knockout occurs whenever the gene product is not expressed. In general, DNA sequences could be excised, replaced, or have DNA sequences added to them. In addition to deleting a target gene, frameshift mutations or DNA sequence insertions that cause the target gene to lose its function also yield gene knockouts.

Alpha-1-Antitrypsin and Monocyte-Activating Polypeptide I Knockouts

26. There was a reasonable expectation that the use of knockout technology as available in 2003 with the alpha-1-antitrypsin (*AAT*) and endothelial monocyte-activating polypeptide I (*EMAP*) genes would have been successful. While gene knockout with any gene is relatively straightforward, knocking out small genes is even simpler. Knockout strategies tend to be simpler for genes having protein coding exons spanning fewer than 20 kb because all protein coding exons may be deleted. By contrast, for larger genes, either a portion of the exons is deleted or the gene is mutated in some other manner. In some situations, a mutated gene may produce a gene product, and thus knockout may not be totally effective. When the entire gene is deleted, however, no gene products will be produced. Because *AAT* and *EMAP* are relatively small (<20 kb), those entire genes could be knocked out. The likelihood of successfully knocking out *AAT* and *EMAP* is high, especially considering the relatively small size of those genes. While deleting the entire *AAT* and *EMAP* genes would have been most preferable, mutations other than total deletion would also have been a straightforward and acceptable means of gene knockout.

27. The success of RNA antisense experiments with *AAT* and *EMAP* suggests that normal expression of AAT and EMAP is not required for cell survival. Because the use of antisense methods did not prove lethal to cells, one would not expect *AAT* or *EMAP* knockout to be lethal.

Furthermore, null *AAT* homozygote humans are known to survive despite AAT deficiency. (see, Cox *et al.* (1988) Am. Rev. Respir. Dis. 137: 371-5, Abstract; attached as Exhibit C).

28. There is a high likelihood that *AAT* and *EMAP* knockout would produce the same phenotypes as found in our corresponding antisense studies with those genes. While it is generally difficult to predict knockout phenotype for a given gene, antisense phenotype generally correlates with knockout phenotype. In light of our antisense studies, the *AAT* and *EMAP* knockout phenotypes would have been fairly predictable as of 2003.

Evidence of AAT and EMAP Down-Regulation by Antisense

29. Exhibit D is attached hereto and provides evidence for the inhibition of EMAP using antisense vectors in CHO cells. Because we were unsuccessful in generating anti-EMAP antiserum, Exhibit D shows proof of concept knockdown of an epitope tagged (V5) version of EMAP by titrating increasing amounts of EMAP antisense vector. The left side of the figure is a Western blot of transfected lysates blotted with an anti-V5 antibody.. As increasing amounts of antisense are introduced, the level of EMAP expression decreases. The right side of the figure is a Coomassie stained blot, showing equal protein loading.

30. Exhibit E is attached hereto and demonstrates that inhibition of both AAT and EMAP expression correlates with an increase in antibody production. The left side of the figure is an ELISA showing the increase in antibody titer caused by the introduction of AAT/EMAP

antisense constructs. The right side of the figure is a Western blot of supernatants from the cells (AAT is secreted), showing the decreased level of AAT expression. EMAP levels could not be analyzed due to the absence of a satisfactory anti-EMAP antibody.

No Expectation of AAT or EMAP Functional Redundancy

31. The superfamily of serine protease inhibitors (SERPINs) plays a key role in controlling the activity of proteinases in diverse biological processes (Seixas *et al.* (2007) Mol. Biol. Evol. 24: 587-598; attached as Exhibit F). Alpha1-antitrypsin (AAT), the most studied member of this family, is encoded by a gene located within the SERPIN gene cluster. While the gene family shares the general activity of being protease inhibitors, it is widely believed the sole function of AAT is the regulation of serine protease activity. Knockout of alpha-1-antitrypsin would most likely not be counteracted by SERPIN family redundancy, including redundancy among the *AAT* genes, as evidenced by alpha-1-antitrypsin deficiency in humans. The lack of AAT secretion causes severe serum deficiency predisposing to chronic lung disease. If there were functional redundancy, this effect would not be observed.

32. EMAP, or S100A11 is an EF-hand motif containing protein. S100 proteins comprise a multigene family and undergo conformational changes in response to calcium binding, permitting dimer formation and altering the activity of specific target proteins. They can regulate cell differentiation, cell cycle progression, energy metabolism, kinase activity, and cytoskeletal membrane interactions. (Ruse *et al.* (2001) Biochem. 40: 3167-3173; attached as

Exhibit G). While this family of proteins shares a similar mechanism of activation, data suggests that function of EMAP may stand apart from the other protein family members. S100A11 was found to have a localization distinct from other S100 proteins examined, being mostly localized to the nucleus (Inada *et al.* (1999) Biochem. Biophys. Res. Commun. 263: 135-138; attached as Exhibit H). Therefore, this suggests knockout of *EMAP* is an isolated event, with little chance of functional redundancy.

33. The Examiner correctly notes that gene knockout may, in some instances, lead to an avalanche of compensatory processes and resulting secondary phenotypical changes. I disagree with the Examiner's assertion, however, that the possibility of developing secondary phenotypical changes would be problematic when producing *AAT* and *EMAP* knockouts for enhancing antibody expression in antibody producing cells, as there is little possibility of functional redundancy, as described above. One of skill in the art would expect any relevant secondary phenotypical changes occurring with gene knockout to also be present in antisense studies employing the same genes. The existence of secondary phenotypical changes may be problematic when ascertaining a particular protein's function because it is desirable to attribute all phenotype changes to a particular gene. The present invention does not, however, seek to determine the function of AAT or EMAP. Rather, the goal of knocking out the *AAT* and *EMAP* genes is to produce high titer antibody producing cells. Because the invention does not specifically seek to ascertain the function of AAT or EMAP and there is little possibility of functional redundancy, any secondary phenotypical changes incidental to the production of high

titer antibody producing cells are irrelevant. Because antisense studies resulted in the desired phenotype, *i.e.*, high titer antibody producing cells, any secondary phenotype changes resulting from *AAT* or *EMAP* knockout are expected to be present in both antisense and gene knockout cells and are not likely to adversely affect the desired phenotype. Indeed, no adverse effects on cell viability or other phenotypes were observed in the antisense system described in the application. Therefore, the Examiner's assertion that the *AAT* and *EMAP* knockout phenotypes would be unpredictable is not correct.

Starting Materials are Adequately Disclosed by the Application

34. Several resources existing in 2003 would have made the production of *AAT* and *EMAP* knockout cells straightforward. All required experimental conditions are described in detail in the literature. Much of the work that would be required for *AAT* and *EMAP* knockout was already performed when we produced antisense transcript expression vectors. Extension of those findings to produce *AAT* and *EMAP* knockouts would have been straightforward to an ordinary scientist in the field at the time of this application's filing. Moreover, producing a gene knockout tends to be more straightforward than performing antisense with the same gene.

35. In 2003, an ordinary scientist in the field would have found that the starting materials for *AAT* and *EMAP* knockouts were adequately disclosed in the application such that knockout cell lines could have been produced without undue or excess experimentation. One seeking to knock out *AAT* and *EMAP* in 2003 could have employed the methods described above with a

reasonable expectation of success. Briefly, it would first be necessary to ascertain the gene of interest's sequence. The sequences of *AAT* and *EMAP* from various species were known and publicly available in various databases as of 2003, as were those genes' flanking sequences. Thus, a determination of the relevant sequences for *AAT* and *EMAP* genes, even in the case of multiple isoforms, for a species of interest can readily be determined using techniques that are routinely practiced in the art. With these sequences in mind, targeting vectors for *AAT* and *EMAP* could have been designed to include regions homologous to the target genes and their flanking sequences, positive selection markers, negative selection markers, sites of targeting vector linearization (assuming the use of bacterial plasmids), and restriction enzyme sites for screening. Commercial vectors were available for this purpose. Alternatively, a suitable vector could have been produced de novo, using nothing more than routine recombinant DNA techniques.

36. Additionally, the variety of eukaryotic drug selection markers makes the matter of knocking out both *AAT* and *EMAP* genes in a cell line completely feasible. In such a case sequential targeting would occur in which one gene locus was knocked out using a suitable drug selection marker such as neomycin phosphotransferase (*neo*). After confirmation of the knockout, the line could then be used for a second round of homologous recombination to knock out the second locus using a selection marker such as hygromycin B phosphotransferase (*hyg*). Sequential knockdown was routinely practiced as of 2003, and in fact, the literature is replete with examples of dual and even triple knockdown cells and organisms such as mice.

37. Once a proper targeting vector was designed, the production of bacterial plasmids would have been straightforward. Plasmids could have been introduced into the antibody producing cells, and cells could have been screened for proper incorporation of the targeting vector. Because each of the necessary starting materials for production of *AAT* and *EMAP* knockouts were properly disclosed by the application and readily available, I disagree with the Examiner's assertion that the application does not adequately disclose the starting materials for the production of *AAT* and *EMAP* knockout cells.

Conclusions

38. The state of gene knockout technology as of this application's filing date in 2003 was such that an ordinary scientist in the field could have produced *AAT* and *EMAP* knockout cell lines based on this application's disclosure. Thousands of journal articles and numerous textbooks describe the methods in sufficient detail for there to have been a reasonable expectation of success in producing those knockout cell lines. Antisense studies performed with *AAT* and *EMAP* suggest that knockout of these genes would have produced viable cells. Moreover, one would expect the *AAT* and *EMAP* antisense phenotypes to be the same as the *AAT* and *EMAP* knockout phenotypes. Therefore, the *AAT* and *EMAP* knockout phenotypes are predictable given the application's disclosure, and the application adequately discloses the necessary starting materials for *AAT* and *EMAP* knockout production.

Date: August 6, 2007

J. Bradford Kline, Ph.D.

Attachments
Exhibits A-H

Exhibit "A"

J. Bradford Kline, Ph.D.

3210 Sunset Ave
Norristown, PA 19403
Home: (610) 584-9502 Work: (610) 423-6123
email: bkline@morphotek.com

SUMMARY Senior research scientist with proven leadership skills in the development and application of new technologies for enhancing throughput and efficiency in drug discovery and therapeutic development.

EXPERIENCE

June 1, 2007 to present **Assistant Director, Genomics/Ab Core**
Morphotek Inc. Exton, PA

- Report directly to the Senior Vice President for Research
- Cloning and expression optimization of recombinant antibodies for production scaleup

June 1, 2003 to June 1, 2007 **Group Leader**
Morphotek Inc. Exton, PA

- Coordinate and lead company-wide screening campaigns utilizing Morphotek's DIRECT-LINE™ technology for enhanced antibody producing cell lines.
    - Optimizes proprietary therapeutic antibodies before integration into clinical trials.
        - Enhanced productivity
        - Improved growth rates
- Using Morphotek's proprietary technology, lead contract screening projects to enhance a variety of cell line traits. Collaborations with companies including:
    - Wyeth
    - Abgenix
    - NovoNordisk
    - GlaxoSmithKline

June 1, 2001 to June 1, 2003 **Senior Scientist**
Morphotek Inc. Philadelphia, PA

- Strategic planning and execution of basic research in the newly established Mammalian Genetics department.
- Led optimization of the *in vivo* MORPHOGENICS™ process to generate models for drug discovery and development including:
    - Eukaryotic cell lines
    - Prokaryotic cell lines
    - Transgenic mice

EDUCATION

1996 to 2001 University of Pennsylvania Philadelphia, PA
Post-doctoral Fellow, Department of Pathology and Lab Medicine
Area of Research: Molecular Endocrinology- Prolactin receptor function in signal transduction.

1992 – 1996	University of Miami Degree: Ph.D. Dissertation Research: Streptococcal virulence factors.	Miami, FL
1985 – 1989	University of Virginia Degree: B.A., Major: Biology	Charlottesville, VA

PATENTS

1. Clevenger C.V., **Kline J.B.** US Patent 6,867,187: Composition and method for modulating somatolactogenic function, March 15, 2005.
2. Nicolaides N., **Kline J.B.** US Patent 10/624,631: Methods for generating enhanced antibody-producing cell lines with improved growth characteristics.

PUBLICATIONS

1. Li, J., T. Sai, M. Berger, Q. Chao, D. Davidson, G. Deshmukh, B. Drozdowski, W. Ebel, S. Harley, M. Henry, S. Jacob, B. Kline, E. Lazo, F. Rotella, E. Routhier, K. Rudolph, J. Sage, P. Simon, J. Yao, Y. Zhou, M. Kavuru, T. Bonfield, M. Thomassen, P. Sass, N. Nicolaides, L. Grasso. 2006. "Human antibodies for immunotherapy development generated via a human B cell hybridoma technology." *PNAS.* **103**:3557-3562.
2. Nicolaides, N.C., W. Ebel, **B. Kline**, Q. Chao, E. Routhier, P. Sass, and L. Grasso. 2005. "Morphogenics as a Tool for Target Discovery and Drug Development." *Annuals of the New York Academy of Sciences.* **1059:**86-96.
3. Grasso, L., **J.B. Kline**, Q. Chao, E. L. Routhier, W. Ebel, P. M. Sass, and N. C. Nicolaides. 2004. "Enhancing therapeutic antibodies and titer yields of mammalian cell lines." *BioProcess.* **2:**58-64.
4. **Kline, J. B.** and C.V. Clevenger. 2002. " Characterization of a Novel and Functional Human Prolactin Receptor Isoform (ΔS1PRLr) Containing Only One Extracellular Fibronectin-like Domain." *Molecular Endocrinology.* **16:**2310-22.
5. **Kline, J. B.** and C.V. Clevenger. 2001. "Identification and Characterization of the Prolactin-binding Protein in Human Serum and Milk.". *The Journal of Biological Chemistry.* **6:**24760-24766.
6. **Kline, J. B.,** D.J. Moore and C.V. Clevenger. 2001. "Activation and Association of the Tec Tyrosine Kinase with the Human Prolactin Receptor: Mapping of a Tec/Vav – Receptor Binding Site." *Molecular Endocrinology.* **15:**832-41.
7. Clevenger, C.V. and **J.B. Kline.** 2001. "Prolactin Receptor Signal Transduction." *Lupus.* **10:**706-718.
8. Clevenger, C.V., M.A. Rycyzyn, F. Syed, and **J.B. Kline.** 2000. Prolactin receptor signal transduction. In: *Prolactin.* Horseman, N.D., ed., Kluwer Publishers, New York, NY.
9. Thompson, C.J., C.V.Clevenger, **J.B. Kline**, and S. Ho. 2000. "Identification of a Novel Prolactin Receptor Isoform in Normal and Malignant Rat Tissue." *(Submitted).*
10. **Kline, J. B.,** H.A. Roehrs, and C.V. Clevenger. 1999. "Functional Characterization of the Intermediate Isoform of the Human Prolactin Receptor." *The Journal of Biological Chemistry.* **274**:35461-35468.
11. Papageorgiou, A.C., C.M. Collins, D.M. Gutman, **J.B. Kline**, S.M. O'Brien, H.S. Tranter, and K.R. Acharya. 1999. "Structural basis for the recognition of superantiger streptococcal pyrogenic exotoxin A (SpeA1) by MHC class II molecules and T-cell receptors." *The EMBO Journal.* **18**:9-21.
12. Clevenger, C.V., D.O. Freier, and **J.B. Kline**. 1998. "Prolactin receptor signal transduction in cells of the immune system." *Journal of Endocrinology.* **157**:187-197.
13. **Kline, J. B.** and C.M. Collins. 1997. "Interactions of the Streptococcal Superantigen SpeA with the Human T Cell Receptor." *Molecular Microbiology.* **24**:191-202.

14. **Kline, J.B.**, S. Xu, A.L. Bisno, and C.M. Collins. 1996. "Identification of a Fibronectin-Binding Protein (GfbA) in Pathogenic Group G Streptococci." *Infection and Immunity*. **64**:2122-2129.

15. **Kline, J.B.** and C.M. Collins. 1996. "Analysis of the Superantigenic Activity of Mutant and Allelic Forms of Streptococcal Pyrogenic Exotoxin A." *Infection and Immunity*. **62**:861-869.

Exhibit "B"

PROTOCOL

Contemporary Gene Targeting Strategies for the Novice

*Siew-Sim Cheah and Richard R. Behringer**

Abstract

Gene targeting in mouse embryonic stem (ES) cells is a fundamental methodology for generating mice with precise genetic modifications. Although there are many complex gene targeting strategies for creating a variety of diverse mutations in mice, most investigators initially choose to generate a null allele. Here we provide a guide for the novice to generate a null allele for a protein coding gene using a fundamental gene targeting strategy. Ultimately, a well considered gene targeting strategy saves significant amounts of time, money, and research animal lives. The straightforward strategy presented here bypasses many of the pitfalls associated with gene knockouts generated by novices. This guide also serves as a foundation for subsequently designing more complex gene targeting strategies.

Index Entries: Gene targeting; embryonic stem cells; Cre/loxP; knockout; homologous recombination.

1. Introduction

Gene targeting in mouse embryonic stem (ES) cells has become a routine methodology to study gene function in vivo *(1)*. Indeed, gene targeting core facilities have been established at numerous institutions to facilitate the generation of targeted ES cell lines and the production of mouse chimeras. Although these facilities have centralized many of the unique skills and reagents that are required to generate "knock out" mice, the investigator must still generate a thoughtful strategy to mutate the gene of interest and to subsequently characterize the resulting mutant mice.

For the novice, the details that are required to design the optimal strategy to generate a targeted mutation are usually not considered because there are so many targeting strategies that are possible *(2)*. The novice often becomes dazzled by these options and chooses strategies that are highly complex and therefore likely to fail. In addition, beginners often become so excited about generating their first knock-out mutation that they rush to generate their gene targeting vector without carefully considering all of the consequences. Inevitably, these short cuts come back to haunt them, resulting in unnecessary and costly delays and many times starting over to retarget the locus using a different strategy. By investing the time in a thoughtful design of a gene targeting strategy, one actually saves valuable time and money at subsequent stages of the experiment. Here we provide a fundamental strategy to generate a null allele for a protein coding gene by gene targeting in ES cells.

2. General Gene Targeting Strategy

Typically, the first desired mutation to generate in a protein coding gene is a null allele. There are generally two situations to consider. In the first situation the protein coding exons of your gene span a relatively small genomic distance (20 kb or less). In the second situation the protein coding exons of your gene span a relatively large distance (greater than 20 kb). For a small gene, the gene targeting strategy is simple, delete all of the pro-

Author to whom all correspondence and reprint requests should be addressed: Siew-Sim Cheah and Richard R. Behringer, Department of Molecular Genetics, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, *E-mail: rrb@notes.mdacc.tmc.edu

Molecular Biotechnology ©2001 Humana Press Inc. All rights of any nature whatsoever reserved. 1073–6085/2001/19:3/297–304/$12.00

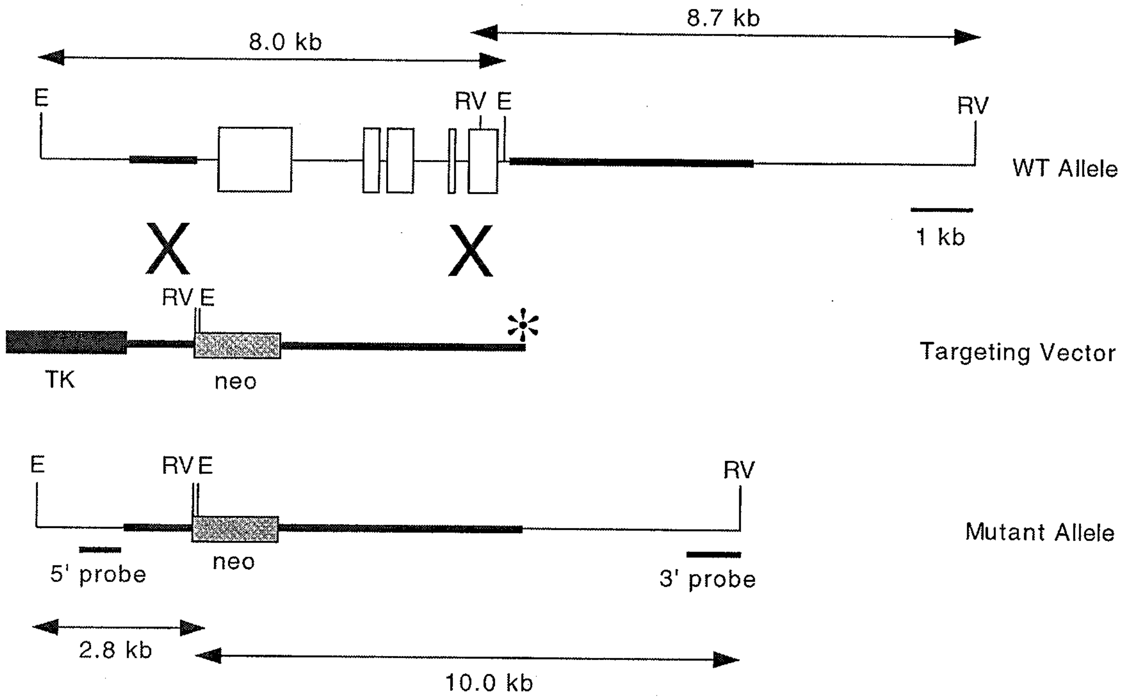


Fig. 1. Deletion of all protein coding exons for a small gene. The example shown is for the generation of a null allele for *Lim1* *(5)*. All of the protein coding exons (open boxes) were replaced by a neomycin (*neo*) cassette to ensure the generation of a null allele. In this case, the total amount of homology (5.5 kb) was divided as a 1.2-kb 5' arm and a 4.3-kb 3' arm (thick lines). A thymidine kinase (tk) cassette was placed adjacent to the 5' arm of homology. The site of vector linearization is indicated (*). In this situation, 5' and 3' external probes were used to identify targeting events. The sizes of the wild-type (WT) and mutant bands for Southern analysis are indicated. E, *Eco*RI; RV, *Eco*RV.

tein coding exons *(3–5)*. Using this strategy, you guarantee the generation of a null allele because it is impossible to generate a protein product from the targeted locus **(Fig. 1)**. In this case, once you have verified that the DNA coding sequences are deleted, it is not necessary to perform mRNA or protein analyses for the deleted gene. Thus, your time can be allocated toward phenotypic characterization studies. If coding sequences remain, mRNA and protein studies of the mutated gene can sometimes be difficult and confusing *(6,7)*. In many situations, biochemical or molecular biological studies have identified domains that are essential for the activity of a specific protein in vitro. This leads some to only delete specific domains because they believe by doing so any partial protein that is generated would have no function. This is a fundamental error in logic because partial protein products may have unknown functions. For a larger protein coding gene, this logic does have some validity due to the constraints imposed by targeting a large gene. But for small genes, we strongly urge that all protein coding exons be deleted.

Mutating a larger gene requires more thought because simple gene targeting strategies are less efficient for generating very large deletions (>20 kb) ***(8–10)*** and may require sophisticated gene targeting skills that the novice has yet to develop *(11,12)*. In addition, large deletions may remove other genes or regulatory sequences of neighboring genes that reside within large genes. The general strategy we suggest is to generate up to a 20 kb deletion to remove as many of the protein coding exons as possible including the exon containing the translation initiation codon. We also suggest that the transcription initiation site, if known, also be included

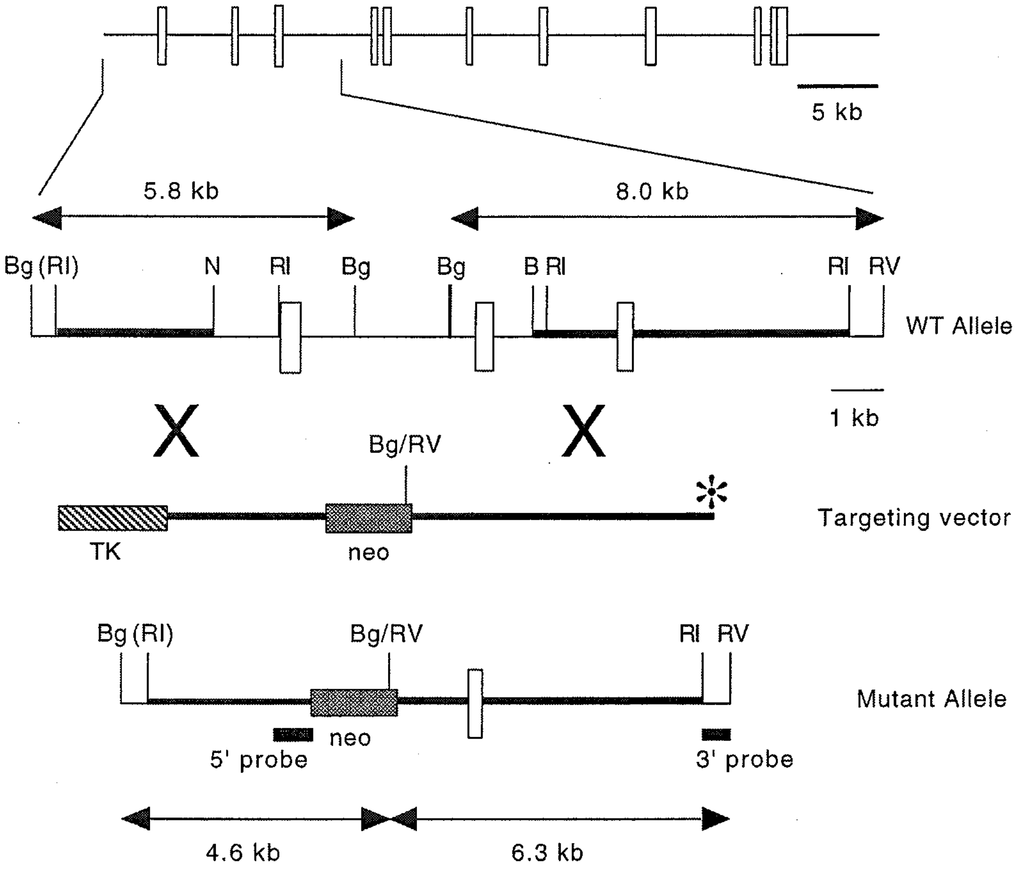


Fig. 2. Deletion of the initial protein coding exons for a large gene. The example shown is for the *Bmpr* locus *(13)*. The coding exons (open boxes) of this gene span 38 kb (top). Therefore, the first two coding exons were deleted. This region (6.3 kb) was replaced by a *neo* cassette. A total of 9.0 kb of homology was divided as a 3.0-kb 5' arm and a 6.0-kb 3' arm. A tk cassette was placed on the 5' arm of homology. The site of vector linearization is indicated (*). A 3' external probe was used for the initial identification of homologous recombinants. Once these clones were identified they were expanded for subsequent analysis with an internal probe to characterize the structure of the 5' recombination event. The sizes of the WT and mutant bands for Southern analysis are indicated. Bg, *Bgl*I; N, *Nco*I; (RI), *Eco*RI from phage multicloning site; RV, *Eco*RV.

in the targeted deletion (**Fig. 2**). If you design a deletion that removes one or more internal protein coding exons then check that the remaining exons are not in frame to generate a partial protein product. Although the above suggestions do not guarantee the generation of a null allele, it increases its likelihood *(13,14)*.

The general strategy we describe uses a replacement gene targeting vector *(15)*. The standard features of a replacement vector are a plasmid backbone containing a positive selection cassette placed between two regions of chromosomal homology and a negative selection cassette adjacent to one of the homologous arms (**Fig. 3**). A positive/negative selection scheme is employed to enrich for homologous recombinants *(16)*. For a replacement strategy, the gene targeting vector is linearized outside of the regions of homology before introduction into mouse ES cells by electroporation. Drug resistant ES cell colonies are picked into 96-well tissue culture plates for expansion and analysis by Southern blot *(17)*. Once the correctly targeted ES cell clones have been identified they are used for the generation of mouse chimeras *(18)* or for in vitro differentiation *(19)* or the generation of teratomas *(20)*.

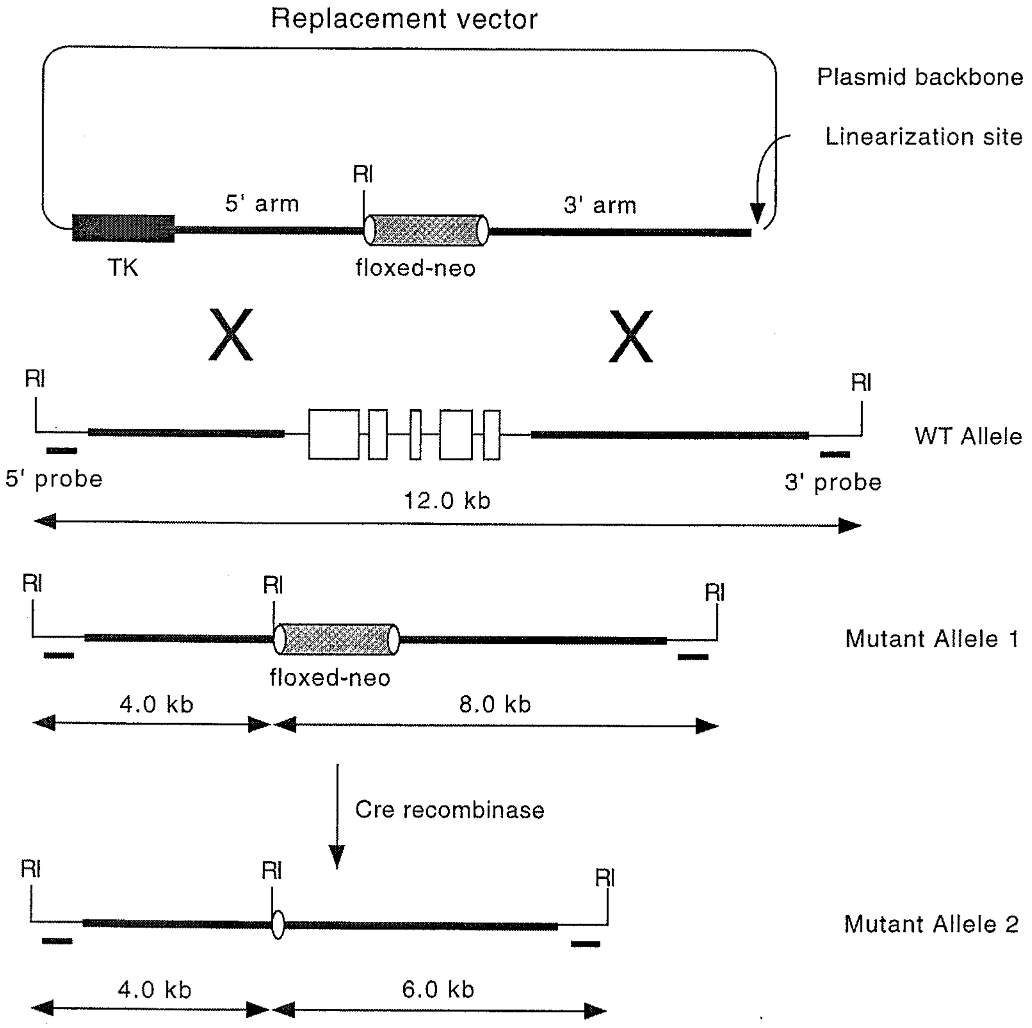


Fig. 3. General design of a replacement gene targeting vector. The floxed *neo* and tk cassettes can be placed in either forward or reverse orientation relative to the orientation of transcription for the gene to be targeting. Mutant allele 1 represents the targeted locus with *neo* and mutant allele 2 represents the targeted locus that has had the *neo* cassette removed by cre recombinase. The sizes of the WT and mutant bands for Southern analysis are indicated.

3. Method

3.1. Isolate Multiple Clones for Your Gene from a Mouse Genomic Library

In constructing your targeting vector, it is important to use genomic DNA that is isogenic to the mouse ES cell line that will be used for the gene targeting experiments to facilitate the maximum frequency of homologous recombination events *(21)*. For historical reasons, the majority of mouse ES cell lines are derived from 129 inbred mouse strains *(22)*. It is essential to note that there are many substrains of 129 that can be quite genetically diverse, especially the 129/SvJ substrain *(23)*. Therefore, it is important to determine the precise strain and substrain of mouse used to generate the ES cell line that you will use in your gene targeting experiments and screen a genomic library from that same strain/substrain. One currently very popular mouse ES cell line called R1 is derived from a (129/Sv x 129/SvJ)F$_1$ embryo *(24)*. 129 strain mouse genomic libraries are commercially available from numerous sources (Stratagene, La Jolla, CA; Genome Systems, St. Louis, MO; Research Genetics, Huntsville, AL).

It is important to isolate multiple genomic clones for your gene especially if it is a large gene. This will provide more genomic sequences and thus more options in designing a targeting vector. In addition, be aware that genomic libraries may contain single clones that contain DNA fragments from different regions of the genome. The coligated genomic inserts in these clones can eliminate all chances of generating or recognizing a targeted mutation. Generate a detailed restriction map for each genomic clone and verify that this map truly matches the chromosomal locus by Southern blot hybridization of isogenic genomic DNA derived from cells or tissues. There have been many instances when correct gene targeting had been obtained but results of the Southern analysis were not as predicted because of an erroneous restriction map.

3.2. Replacement Vector Design

Once you have obtained and characterized genomic clones for the gene of interest, you must design your gene targeting strategy "on paper." We urge beginners to develop a gene targeting strategy that employs a Southern blot analysis to identify targeting events. The 96-well tissue culture plate method for the rapid screening of ES cell clones by Southern analysis is exceedingly easy for beginners to master very quickly *(17)*. We do not recommend a PCR genotyping strategy, because it imposes constraints on your gene targeting vector design. In addition, we have found that PCR genotyping is more problematic for beginners than Southern blot analysis. Furthermore, PCR results must always be confirmed by Southern analysis. In your design, identify genomic regions that can serve as Southern blot probes to recognize targeting events. Do not generate your targeting vector until your potential probes have been verified on Southern blots of isogenic genomic DNA. Once the probes have been verified, then construct your gene targeting vector as shown in **Fig. 3**. The details of each aspect of the replacement vector are discussed below.

3.2.1. Choice of Homologous Regions for the Targeting Vector

The total amount of chromosomal homology should be between 5 and 8 kb *(15)*. Smaller amounts of homology may reduce the targeting frequency and significantly larger amounts of homology may become unwieldy for targeting vector construction and homologous recombinant identification. The amount of homologous sequences for each arm of homology in the vector should be about half of the total amount of homology. For example, a vector with a total of 6 kb of homology may have two arms of homologous sequence that are approx 3 kb each.

Each arm of homology in the targeting vector should flank the protein coding exons and other sequences that you wish to delete. Remember that neighboring genes may lie very close to your gene of interest *(25)*. Therefore, try not to delete beyond the known sequences of your gene.

3.2.2. Positive Selection Cassette

We suggest that you use a neomycin phosphotransferase (*neo*) gene expression cassette for selection with G418 that has the mouse phosphoglycerate kinase (PGK) promoter and the polyadenylation signal from the bovine growth hormone (bpA) gene (*PGKneobpA*) because it is more efficient than other *neo* expression cassettes that are typically used in gene targeting experiments *(26)*. Although there are other positive selectable markers that can be used, for example, hygromycin, puromycin, and so on, we suggest that the beginner use neomycin because it is routinely used by most laboratories and most readily available feeder cell lines, on which ES cells grow, are G418 resistant. The *neo* cassette can be placed in either forward or reverse orientation relative to the direction of transcription of your gene.

The presence of a selectable marker cassette with its exogenous promoter has been documented to alter gene expression at targeted loci *(27)*. Therefore, the *neo* cassette in the strategy we outline should be flanked by loxP sequences (floxed) to provide the option of subsequently removing the *neo* cassette with Cre recombinase *(28)*. You lose nothing by cloning a floxed-*neo* expression cassette into your gene targeting vector and gain many potential benefits. The option to remove the *neo* cassette using simple methods can potentially save effort and the costs of regenerating another mutant should any question arise about the altered transcription of your gene or neighboring genes *(29,30)*. The floxed-*neo* cassette strategy also provides the opportunity to remove the selectable marker cassette from the targeted locus in vitro for subsequent retargeting of the remaining wild-type allele with the original targeting vector and drug selection to generate homozygous mutant ES cell lines, a process called "marker recycling" *(31)*. The floxed-*neo* cassette strategy also allows one to easily generate two different alleles for your gene of interest, one with *neo* and one without, that can be distinguished by Southern blot or PCR. The availability of two different alleles can be exploited for potential chimera experiments *(32,33)*. The floxed-*neo* cassette can be removed in ES cells by transient Cre expression in vitro *(31)* by the injection of a Cre expression vector into the pronuclei of eggs containing the floxed gene *(34)*, or by simply crossing mice carrying the floxed gene with Cre expressing transgenic mice *(35)*.

3.2.3. Negative Selection Cassette

We suggest that you use a herpes simplex virus thymidine kinase gene expression cassette for negative selection with gancyclovir or FIAU (1-2-deoxy-2-fluroro-1-β-D-arabinofuranosyl-5-iodouracil). The MC1TKpA vector works very well in our hands with FIAU, providing a 5- to 10-fold enrichment for targeting events *(36)*. An alternate and convenient negative selection cassette is one that expresses the diphtheria toxin A-chain (DT-A) gene *(37)*. The advantage of the DT-A strategy is that it is not necessary to add any drug selection to the culture medium because expression of DT-A in cells is toxic.

The beginner should note that even after positive-negative drug selection, many of the resulting ES cell colonies are nontargeted (random integration) events. This occurs because the negative selection cassette used in the gene targeting vector is not expressed, possibly due to damage or integration into chromosomal regions that inhibit its expression. The negative selection cassette can be placed in either forward or reverse orientation relative to the direction of transcription of your gene.

3.2.4. Site of Targeting Vector Linerization

Gene targeting vectors are routinely introduced into ES cells by electroporation in a linearized form. Thus, in the design of your targeting vector, a unique restriction enzyme site must be present as a site for linearization. For a replacement gene targeting vector, the linearization site must be outside of the regions of homology. Typically, the linearization site is located at the junction of one of the homologous arms and the plasmid backbone. We suggest that the linearization site should not be at the junction between the negative selection cassette and the plasmid backbone **(Fig. 3)** to reduce the chances of the negative selection cassette from being degraded by nucleases after introduction into cells, causing an increase in the background of nontargeted clones.

3.2.5. Restriction Enzyme Sites for Southern Analysis

To identify homologous recombinants, Southern analysis using a diagnostic restriction enzyme digest with a probe that is external to the regions of homology (external probe) included in the targeting vector must be performed. Therefore, it is important to consider restriction enzyme sites that can be used to differentiate between targeted and nontargeted events during the construction of the targeting vector. A unique restriction enzyme site should be introduced by the positive selection cassette. Remember that this restriction enzyme site should be located outside of the floxed *neo* expression cassette so that after Cre expression the diagnostic restriction enzyme site will still remain. In this way, a diagnostic digest will yield a smaller DNA fragment for the mutant allele in comparison to the wild-type allele that can be recognized by the external probe. This is essential because partial restriction enzyme digestions can occur in the cruder DNA samples that are prepared using the 96-well tissue culture plate method for the rapid screening of recombinant ES cells. Thus, if the mutant allele yields a DNA fragment that is larger than the wild-type allele, a partial restriction enzyme digest can cause difficulties during the Southern analysis. We suggest choosing restriction enzymes that work efficiently on these cruder DNA preparations for the most consistent results for genomic Southerns. The restriction enzymes that have worked well for us include *Bam*HI, *Bgl*I, *Bgl*II, *Eco*RI, *Eco*RV, *Pst*I, and *Sst*I whereas *Kpn*I, *Xba*I, *Xho*I, and *Xmn*I have proven to be problematic.

3.2.6. 5' and 3' External Probes for the Identification of Targeted ES Cell Clones

It is important to use both 5' and 3' external probes that can recognize the structure of the gene targeting events on the 5' and 3' arms of homology. It is not uncommon for correct targeting to be obtained on one arm of homology but not the other. If both 5' and 3' external probes cannot be found, the initial targeting event can be identified using one external probe that confirms the structure of the targeting event on one side of homology. Once this subset of targeted ES cell clones is identified, then a probe within the region of targeting vector homology (internal probe, for example, *neo*) can be used to verify the structure of the targeting event on the other side of homology. Do not use the PGK promoter sequences of the *neo* expression cassette as an internal probe because it is derived from mouse and will recognize the endogenous mouse PGK gene.

4. Summary

Gene targeting in mouse ES cells is a powerful method for studying gene function in vivo. For the novice this combination of molecular biology, specialized tissue culture cell lines, and mouse reproductive biology can be daunting. We present a straightforward, one might say constrained, guide for novices of gene targeting to generate a null allele in large or small protein coding genes. The method we outline has evolved from years of experience training and advising beginners on this powerful technology. We believe that a good design for a gene targeting strategy ultimately saves time, money, and research animal lives. Once you feel comfortable with a fundamental knockout, then we suggest you consider the new and exciting gene targeting variations that can be used to address important biological questions *(2,28)*.

Acknowledgment

The authors thank Yuji Mishina and Maki Wakamiya for helpful comments on the manuscript.

References

1. Torres, M. (1998) The use of embryonic stem cells for the genetic manipulation of the mouse. *Curr. Top. Dev. Biol.* **36,** 99–114.
2. Bradley, A. and Liu, P. (1996) Target practice in transgenics. *Nature Gen.* **14,** 121–123.
3. Chen, Z.-F. and Behringer, R. R. (1995) *twist* is required in head mesenchyme for cranial neural tube morphogenesis. *Genes Dev.* **9,** 686–699.
4. Rivera-Pérez, J. A., Mallo, M., Gendron-Maguire, M., et al. (1995) *Goosecoid* is not an essential component of the mouse gastrula organizer but is required for craniofacial and rib development. *Development* **121,** 3005–3012.
5. Shawlot, W. and Behringer, R. R. (1995) Requirement for *Lim1* in head-organizer function. *Nature* **374,** 425–430.
6. Moens, C. B., Auerbach, A. B., Conlon, R. A., et al. (1992) Targeted mutation reveals a role for *N-myc* in branching morphogenesis in the embryonic mouse lung. *Genes Dev.* **6,** 691–704.
7. Hasty, P., Bradley, A., Morris, J. H., et al. (1993) Muscle deficiency and neonatal death in mice with a targeted mutation in the myogenin gene. *Nature* **364,** 501–506.
8. Mombaerts, P., Clarke, A. R., Hooper, M. L., et al. (1991) Creation of a large genomic deletion at the T-cell antigen receptor beta-subunit locus in mouse embryonic stem cells by gene targeting. *Proc. Natl. Acad. Sci. USA* **88,** 3084–3087.
9. Zhang, H., Hasty, P., and Bradley, A. (1994) Targeting frequency for deletion vectors in embryonic stem cells. *Mol. Cell. Biol.* **14,** 2404–2410.
10. Tsuda, H., Maynard-Currie, C. E., Reid, L. H., et al. (1997) Inactivation of the mouse HPRT locus by a 203-bp retroposon insertion and a 55-kb gene-targeted deletion: establishment of new HPRT-deficient mouse embryonic stem cell lines. *Genomics* **15,** 413–421.
11. Ramírez-Solis, R., Liu, P., and Bradley, A. (1995) Chromosome engineering in mice. *Nature* **378,** 720–724.
12. Li, Z. W., Stark, G., Gotz, J., et al. (1996) Generation of mice with a 200–kb amyloid precursor protein gene deletion by Cre recombinase-mediated site-specific recombination in embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **93,** 6158–6162.

13. Mishina, Y., Suzuki, A., Ueno, N., et al. (1995) *Bmpr* encodes a type I bone morphogenetic protein receptor that is essential for gastrulation during mouse embryogenesis. *Genes Dev.* **9,** 3027–3037.
14. Mishina, Y., Rey, R., Finegold, M. J., et al. (1996) Genetic analysis of the Mullerian-inhibiting substance signal transduction pathway in mammalian sexual differentiation. *Genes Dev.* **10,** 2577–2587.
15. Hasty, P. and Bradley, A. (1993) Gene targeting vectors for mammalian cells, in *Gene Targeting: A Practical Approach* (Joyner, A. L., ed.) IRL Press , Oxford, pp. 1–31.
16. Mansour, S. L., Thomas, K. R., and Capecchi, M. R. (1988) Disruption of the proto-oncogene *int-2* in mouse embryo-derived stem cells: a general strategy for targeting mutations to non-selectable genes. *Nature* **336,** 348–352.
17. Ramírez-Solis, R., Rivera-Pérez, J., Wallace, J. D., et al. (1992) Genomic DNA microextraction: a method to screen numerous samples. *Analyt. Biochem.* **201,** 331–335.
18. Wood, S. A., Allen, N. D., Rossant, J., et al. (1993) Non-injection methods for the production of embryonic stem cell-embryo chimaeras. *Nature* **365,** 87–89.
19. Robertson, E. J. (1987) Embryo-derived stem cell lines, in *Teratocarcinomas and Embryonic Stem Cells: A Practical Approach* (Robertson, E. J., ed.) IRL Press, Oxford, pp. 71–112.
20. Holdener, B. C., Faust, C., Rosenthal, N. S., et al. (1994) *msd* is required for mesoderm induction in mice. *Development* **120,** 1335–1346.
21. te Riele, H., Maandag, E. R., and Berns, A. (1992) Highly efficient gene targeting in embryonic stem cells through homologous recombination with isogenic DNA constructs. *Proc. Natl. Acad. Sci. USA* **89,** 5128–5132.
22. Papaioannou, V. and Johnson, R. (1993) Production of chimeras and genetically defined offspring from targeted ES cells, in *Gene Targeting: A Practical Approach* (Joyner, A. L., ed.) IRL Press, Oxford, pp. 1–31.
23. Threadgill, D. W., Yee, D., Matin, A., et al. (1997) Genealogy of the 129 inbred strains: 129/SvJ is a contaminated inbred strain. *Mamm. Genome* **8,** 390–393.
24. Nagy, A., Rossant, J., Nagy, R., et al. (1993) Derivation of completely cell culture-derived mice from early-passage embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **90,** 8424–8428.
25. Dresser, D. W., Hacker, A., Lovell-Badge, R., et al. (1995) The genes for a spliceosome protein (SAP62) and the anti-Mullerian hormone (AMH) are contiguous. *Hum. Mol. Genet.* **4,** 1613–1618.
26. Soriano, P., Montgomery, C., Geske, R. and Bradley, A. (1991) Targeted disruption of the *c-src* proto-oncogene leads to osteopetrosis in mice. *Cell* **64,** 693–702.
27. Fiering, S., Epner, E., Robinson, K., et al. (1995) Targeted deletion of 5'HS2 of the murine beta-globin LCR reveals that it is not essential for proper regulation of the beta-globin locus. *Genes Dev.* **9,** 2203–2213.
28. Nagy, A. (2000) Cre recombinase: the universal reagent for genome tailoring. *Genesis* **26,** 99–109.
29. Olson, E. N., Arnold, H. H., Rigby, P. W., et al. (1996) Know your neighbors: three phenotypes in null mutants of the myogenic bHLH gene *MRF4*. *Cell* **85,** 1–4.
30. Kaul, A., Köster, M., Neuhaus, H., and Braun, T. (2000) Myf-5 Revisited: Loss of Early Myotome Formation Does Not Lead to a Rib Phenotype in Homozygous Myf-5 Mutant Mice. *Cell* **102,** 17–19.
31. Abuin, A. and Bradley, A. (1996) Recycling selectable markers in mouse embryonic stem cells. *Mol. Cell. Biol.* **16,** 1851–1856.
32. Quinn, J. C., West, J. D., and Hill, R. E. (1996) Multiple functions for *Pax6* in mouse eye and nasal development. *Genes Dev.* **10,** 435–446.
33. Rivera-Pérez, J. A., Wakamiya, M., and Behringer, R. R. (1999) *Goosecoid* acts cell autonomously for the maintenance of mesenchymal tissues during craniofacial development. *Development.* **126,** 3811–3821.
34. Sunaga, S., Maki, K., Komagata, Y., et al. (1997) Efficient removal of loxP-flanked DNA sequences in a gene-targeted locus by transient expression of Cre recombinase in fertilized eggs. *Mol. Reprod. Dev.* **46,** 109–113.
35. Lasko, M., Sauer, B., Mosinger, B. Jr., et al. (1992) Targeted oncogene activation by site-specific recombination in transgenic mice. *Proc. Natl. Acad. Sci. USA* **89,** 6232–6236.
36. McMahon, A. P. and Bradley, A. (1990) The *Wnt-1 (int-1)* proto-oncogene is required for development of a large region of the mouse brain. *Cell* **62,** 1073–1085.
37. McCarrick, J. W., Parnes, J. R., Seong, R. H., et al. (1993) Positive-negative selection gene targeting with the diphtheria toxin A-chain gene in mouse embryonic stem cells. *Transgenic Res.* **2,** 183–190.

Exhibit "C"

Am Rev Respir Dis. 1988 Feb;137(2):371-5.

Emphysema of early onset associated with a complete deficiency of alpha-1-antitrypsin (null homozygotes).

Cox DW, Levison H.

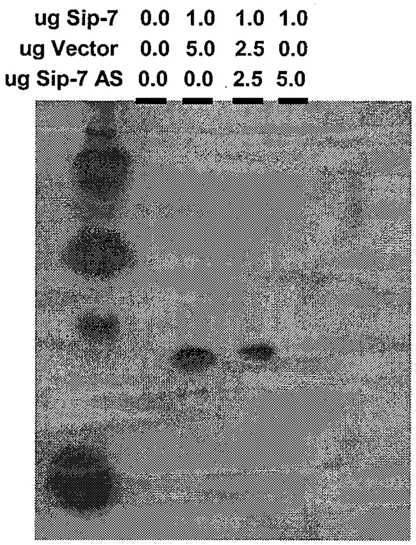
Research Institute, Hospital for Sick Children, Toronto, Ontario, Canada.

We have compared lung function in 3 subjects with no alpha 1-antitrypsin (alpha 1-protease inhibitor) (null homozygotes) with subjects having the typical deficiency, PIZZ. We identified a 31-yr-old woman, presenting with severe obstructive lung disease, who had no detectable plasma alpha 1-antitrypsin, indicating homozygosity for a "null" (or PI*QO) allele of alpha 1-antitrypsin. Two of her sisters have a similar deficiency, one with an onset of symptoms at 17 yr of age. Because of the likelihood that there are a number of different PI*QO alleles, the type in this family has been named null Mattawa (QOmattawa). All 3 homozygotes have shown a marked deterioration of lung function over a 7-yr period of follow-up. In contrast, lung function tests of 6 age-matched nonsmoking subjects with alpha 1-antitrypsin deficiency, PI type ZZ, showed no abnormalities of lung function. The 15 to 20% of the normal plasma concentration of alpha 1-antitrypsin associated with the PI*Z allele appears to provide some protection to the lung in comparison with a complete deficiency state.

Exhibit “D”

Evidence for the Inhibition of Sip-7 using AS vectors in CHO cells

Anti-V5



Coomassie

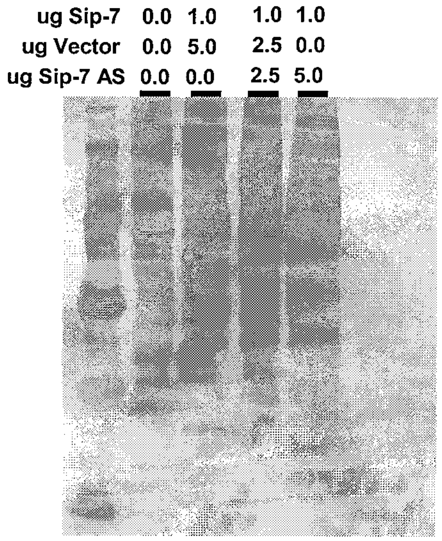
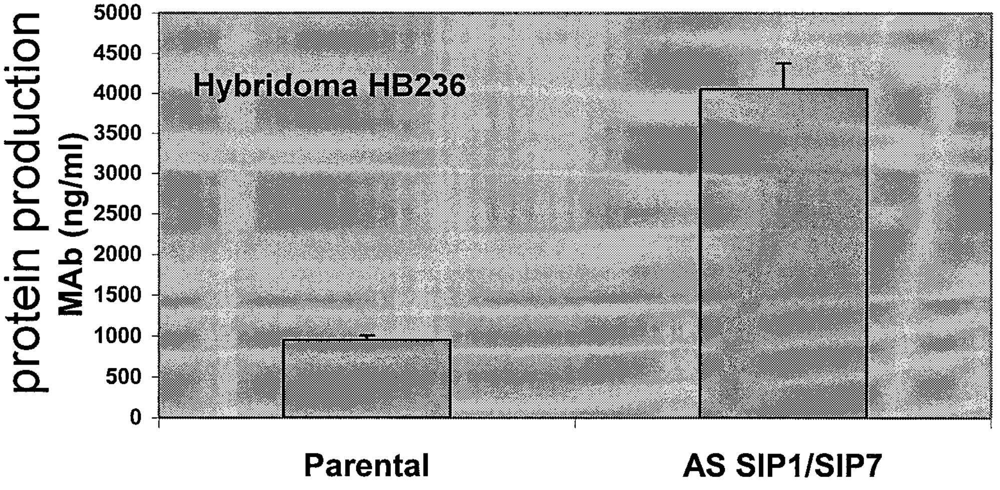


Exhibit “E”

Case Study: using morphogenics for high-titer pathway discovery in hybridomas

validation of Suppressor of Immunoglobulin Production (SIP) genes in increasing titer yields in hybridoma cells

MAb ELISA



Western of SIP1 Expression

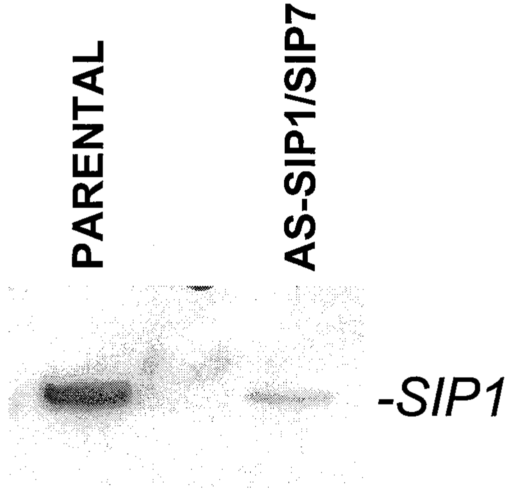


Exhibit "F"

Sequence Diversity at the Proximal 14q32.1 *SERPIN* Subcluster: Evidence for Natural Selection Favoring the Pseudogenization of *SERPINA2*

Susana Seixas,† Gianpaolo Suriano,*‡ Filipa Carvalho,§ Raquel Seruca,*‡ Jorge Rocha,*|| and Anna Di Rienzo†*

*Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal; †Department of Human Genetics, University of Chicago; ‡Faculty of Medicine, University of Porto, Porto, Portugal; §Department of Genetics, Faculty of Medicine, University of Porto, Porto, Portugal; and ||Department of Zoology and Anthropology, Faculty of Sciences, University of Porto, Porto, Portugal

The superfamily of serine protease inhibitors (SERPINs) plays a key role in controlling the activity of proteinases in diverse biological processes. α1-antitrypsin (SERPINA1), the most studied member of this family, is encoded by a gene located within the proximal 14q32.1 *SERPIN* subcluster, together with the highly homologous α1-antitrypsin–like sequence (*SERPINA2*), which was previously proposed to be a pseudogene. Here, we performed a resequencing study encompassing both *SERPINA1* and *SERPINA2* as well as the adjacent gene coding for corticosteroid-binding globulin (*SERPINA6*) in samples from Europe and West Africa. In the African sample, we found that a common haplotype carrying a 2-kb deletion in the *SERPINA2* gene is associated with remarkable long-range homozygozity as if it was quickly driven to high frequency by natural selection acting on an advantageous variant. An analysis of the HapMap Phase I data for the Yoruba sample confirmed that variation in this subcluster carries a strong signal of positive selection. We also show that the *SERPINA2* gene is expressed and probably encodes a functional SERPIN. Finally, comparisons with orthologous sequences in nonhuman primates showed that *SERPINA2* is present in some great apes, but in chimpanzees it was lost by a deletion event independent from that observed in humans. In agreement with the "less is more" hypothesis, we propose that loss of *SERPINA2* is an ongoing process associated with a selective advantage during recent primate evolution, possibly because of a role in fertility or in host–pathogen interactions.

Introduction

Serine protease inhibitors (SERPINs) are a superfamily of highly conserved proteins that are widely distributed among animals, plants, viruses, and bacteria. In vertebrates, these proteins act mainly as protease inhibitors in a number of biological processes such as blood coagulation, complement activation, fibrinolysis, tissue repair, inflammation, and tumor suppression. However, a small fraction of SERPINs have lost their inhibitory ability and developed other specialized roles, such as hormone carriers, chaperones, or storage proteins (Stein and Carrell 1995; Irving et al. 2000, 2002). SERPIN functional diversity and substrate specificity is essentially determined by variation at a reactive center defined by a short stretch of amino acids. This reactive center accumulated substitutions at a greater rate than the rest of the molecule, probably as a result of positive natural selection (Hill and Hastie 1987; Creighton and Darby 1989; Goodwin et al. 1996).

The inhibitory function of SERPINs involves a high level of molecular plasticity, which renders the molecules particularly vulnerable to mutations. Single amino acid changes affecting their mobile reactive loop can cause abnormal protein folding and may lead to the pathogenic polymerization processes that underlie the recently recognized category of conformational diseases (Stein and Carrell 1995; Carrell and Lomas 2002; Lomas and Carrell 2002). This concept is behind the most common serpinopathy, α1-antitrypsin (SERPINA1) deficiency, which is mainly caused by homozygosity for a E342L mutation (the Z allele), affecting 1 in 2,000 to 1 in 7,000 individuals of European ancestry (WHO 1997). The major clinical manifestations of SERPINA1 deficiency are early pulmonary emphysema, due to the unopposed action of neutrophil elastase in the lower respiratory tract and hepatic disease caused by the cytotoxic effect of protein accumulation in the rough endoplasmic reticulum from hepatocytes (Cox 1995; WHO 1997; Needham and Stockley 2004). Furthermore, SERPINA1 has been extensively characterized in human populations as a classical protein polymorphism with 5 additional common alleles: M1Ala213, M1Val213, M2, M3 and S all with normal circulating protein levels and a mildly deficient S variant (Cox 1995; Nukiwa et al. 1996).

In humans, *SERPINA1* is part of a gene cluster, which spans over ~370 kb on chromosome 14q32.1 and includes 10 additional members of the *SERPIN* superfamily. Within this cluster, the *SERPIN* genes are organized into 3 distinct subclusters. The proximal subcluster contains the α1-antitrypsin (*SERPINA1*), α1-antitrypsin–like (*SERPINA2*), corticosteroid-binding globulin (*SERPINA6*), and protein Z inhibitor (*SERPINA10*) genes. The central subcluster harbors the recently characterized vaspin (*SERPINA12*), centerin (*SERPINA9*), and antiproteinase-like 2 (*SERPINA11*) genes. Finally, the distal subcluster contains the kallistatin-like (*KAL-like*), α1-antichymotrypsin (*SERPINA3*), protein C inhibitor (*SERPINA5*), and kallistatin (*SERPINA4*) genes (Namciu et al. 2004; Marsden and Fournier 2005). All these genes have a significant sequence similarity and most share a common gene structure with 1 untranslated exon and 4 coding exons. Accordingly, it has been proposed that they evolved from a common ancestral gene through a series of duplication events (Atchley et al. 2001; van Gent et al. 2003). Except for *SERPINA2* (Bao et al. 1988; Kelsey et al. 1988), all the members of the chromosome 14 cluster were previously shown to be expressed (Namciu et al. 2004; Marsden and Fournier 2005). Although the gene structure and sequence of *SERPINA2* are very similar to those of *SERPINA1*, no promoter could be identified for *SERPINA2* by sequence homology, leading to the proposal

Key words: natural selection, pseudogenization, SERPIN.

E-mail: sseixas@ipatimup.pt.

Mol. Biol. Evol. 24(2):587–598. 2007
doi:10.1093/molbev/msl187
Advance Access publication November 29, 2006

that this is a pseudogene (Bao et al. 1988; Hofker et al. 1988; Marsden and Fournier 2005). However, different studies have yielded contrasting results regarding the extent of *SERPINA2* sequence degeneration. Bao et al. (1988) reported a sequence with preserved RNA splice sites and no premature stop codons, suggesting that the *SERPINA2* gene, if expressed, could encode a new secretory SERPIN with different substrate specificity. In contrast, Hofker et al. (1988) reported a cloned sequence, which bears a critical mutation in the start codon (AT$\underline{\text{G}}$ → AT$\underline{\text{A}}$) and an ~2-kb deletion encompassing exon IV and part of exon V. This deletion was shown to occur at a 30% frequency in a sample from the Dutch population (Hofker et al. 1988).

The characterization of copy number polymorphism in the human genome, including insertions and deletions from a few to hundreds of kilobases, has been the focus of a number of recent surveys (Sebat et al. 2004; Sharp et al. 2005; Tuzun et al. 2005; Conrad et al. 2006; Feuk et al. 2006; Hinds et al. 2006; McCarroll et al. 2006). Although these genome-wide studies have provided a great deal of information on structural variation, little is known about the role of these variants in common disease phenotypes and in human adaptations. In particular, the "less is more" hypothesis posits that loss of function mutations, such as deletions, are an important substrate for natural selection and may be the basis for many evolutionary adaptations (Olson 1999). Hence, evolutionary studies of genomic regions harboring polymorphic deletions, such as the proximal SERPIN 14q32.1 subcluster, will contribute to understanding the significance of this class of common genetic variation.

Here, we performed a resequencing study of the proximal *SERPIN* 14q32.1 subcluster encompassing the 2 most closely related genes, *SERPINA1* and *SERPINA2*, and the adjacent *SERPINA6* gene. By surveying both coding and noncoding regions in 2 ethnically distinct samples from Europe and Africa, we aimed to provide a deeper understanding of the evolution of this subcluster. We found an unusual pattern in the African sample, resulting from the high frequency of a haplotype carrying a 2-kb deletion in the *SERPINA2* gene. This haplotype shows considerable long-range homozygosity across the surveyed region, suggesting that it quickly reached high frequency due to the action of positive natural selection. Furthermore, we show that the nondeleted form of *SERPINA2* is expressed in different human tissues and that the gene is deleted in chimpanzee, but intact in other great apes. These data taken together suggest that recent positive selection favored the loss of *SERPINA2* function and that the pseudogenization process is still ongoing in humans.

Materials and Methods

DNA Samples

Sequence variation was surveyed in DNA samples from unrelated individuals, with known SERPINA1 protein phenotypes, belonging to 2 populations from different backgrounds: Portugal and São Tomé (Gulf of Guinea, West Africa). The island of São Tomé, located 240 km off the coast of Gabon, was peopled at the end of the 15th century by slaves from the nearby coasts of Africa and hence retained the high levels of genetic diversity that are generally observed in the African mainland (Tomas et al. 2002). Population samples of 40 chromosomes of Portuguese origin and 40 chromosomes from the island of São Tomé were selected from larger samples in which SERPINA1 protein polymorphism had been previously studied (Seixas et al. 2001). All samples were collected with informed consent.

Orthologous regions of the proximal *SERPIN* subcluster were also sequenced in 1 chimpanzee (*Pan troglodytes*), 1 gorilla (*Gorilla gorilla*), and 1 orangutan (*Pongo pygmeus*).

Polymerase Chain Reaction and Sequence Determination

Primers for amplification and sequencing were designed on the basis of GenBank (http://www.gdb.org/) sequence entries AL132708 for *SERPINA1* and AL117259 for *SERPINA2* and *SERPINA6*; all nucleotide positions in this article are numbered according to these sequences. To distinguish between chromosomes bearing *SERPINA2* deletion and nondeletion alleles, we performed allele-specific polymerase chain reaction (PCR) using the following reverse primers: 5′-AGT TGG TGC CAT ACA CTA AT-3′ for the deletion and 5′-AGT TGG TGA TGT CAT CCT TG-3′ for the nondeletion.

Sequencing was performed using the ABI BigDye Terminator version 3 cycle sequencing chemistry (Applied Biosystems, Foster City, CA), and electrophoresis analysis was done on an ABI 3100 automated sequencer. All human sequences were assembled and analyzed using the Phred-Phrap-Consed package (Nickerson et al. 1997). All putative polymorphisms and software-derived genotype calls were visually inspected and were individually confirmed using Consed. Details about PCR and sequencing conditions are available from the authors upon request.

Expression Studies by Real-Time PCR

To study the expression of *SERPINA2*, real-time (RT) PCR experiments were performed using cDNA synthesized by reverse transcription assays in total mRNA from liver, leukocytes, and testis samples. Additionally, the expression of the known functional gene *SERPINA1* was studied for comparison. Reverse transcription was performed using the Superscript II RT PCR system (Invitrogen Life Technology, Carlsbad, CA), according to the manufacturer's protocol. Quantitative RT PCR reactions were performed on an ABI Prism 7000 Sequence Detection System with the TaqMan Universal Master Mix (Applied Biosystems), according to the manufacturer's instructions. The expression level of the rRNA gene was used as control. All primers and probes were designed by using Primer Express version 2.0 software. In order to prevent cross hybridization, primers were anchored on a region of low sequence similarity at the exon II–III junction.

Primers for *SERPINA1* were 5′-GTC AAA CAC CTG AAA AAA GAC ACA A-3′ and 5′-CAC TTG CCG TGA AAG GAA ATG-3′, and the probe was 5′-TCT TGC CCT GGT GGA T-3′. For *SERPINA2*, we used the primers 5′-GGA GCT TGA CAG AGA CAC TTT T-3′ and 5′-GGT CTC TCC CAT TTG CCT TTA A-3′ and a 5′-CTC TGG TGA ATT ACA TCT T-3′ probe. For each gene, primers and probe concentrations were 900 nM and 250 nM,

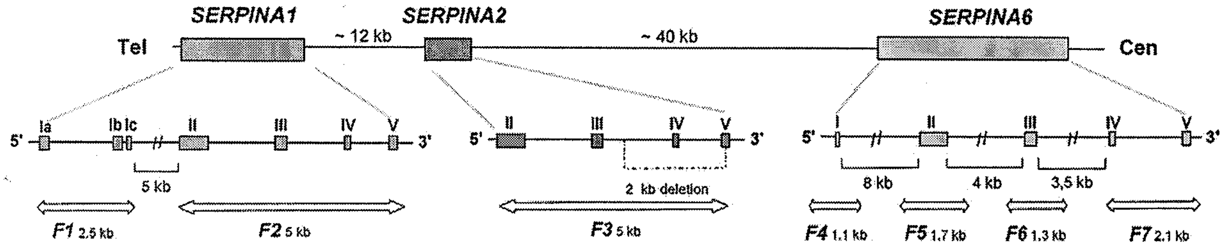


FIG. 1.—Surveyed segments of the 14q32.1 *SERPIN* proximal subcluster. Upper line shows the relative position of the 3 genes on the cluster and lower lines the exon/intron structure of each gene (exons are represented by boxes). Large arrows indicate the regions surveyed for sequence variation (F1–F7). The dotted line represents the 2-kb deletion of *SERPINA2* gene.

respectively. The cycling conditions were as follows: 50 °C for 2 min, 95 °C for 10 min, and 40 cycles of 95 °C for 15 s and 60 °C for 1 min. The expression levels of *SERPINA1* and *SERPINA2* were normalized relative to the expression level of rRNA. All reactions were run in triplicate.

Statistical Analysis

Summary statistics of population genetic variation were calculated using the online applications SLIDER (http://genapps.uchicago.edu/slider/index.html) and MAXDIP (http://genapps.uchicago.edu/labweb/index.html). Haplotypes for *SERPINA2* in deletion-bearing heterozygotes were phased unambiguously by allele-specific PCR and sequencing. The remaining haplotypes were inferred in the total sample by using the program PHASE 2.02. (Stephens et al. 2001; Stephens and Donnelly 2003). $|D'|$ and r^2 were calculated from the inferred haplotypes using the DNAsp program, version 4.0. (Rozas J and Rozas R 1997).

The haplotype test described by Hudson et al. (1994) was performed by simulating 10,000 replicates under neutrality, using estimates of the population recombination and mutation rate parameters calculated for our data using MAXDIP and SLIDER, respectively. We initially performed the test using the best reconstruction of haplotypes provided by the program PHASE. To assess the robustness of results to misspecification of haplotype phase, the tests were subsequently rerun on 100 samples of haplotypes drawn from the posterior distribution provided by PHASE.

The extended haplotype homozygosity statistic (EHH) (Sabeti et al. 2002) was computed using the online tool EHH calculator (http://ihg.gsf.de/cgi-bin/mueller/webehh.pl). To assess the statistical significance of EHH, we performed the long-range haplotype (LRH) test (Sabeti et al. 2002) by comparing the observed values of the relative EHH with theoretical null distributions generated by coalescent simulations with recombination, assuming no selection (Hudson 2002). We simulated 500-kb regions under 4 different demographic models: constant population size, expansion, and bottleneck models described in Voight et al. (2005) and the structured population model used by Sabeti et al. (2002). Comparison of haplotype frequencies versus relative EHH and significance estimation were carried out using the program Sweep 1.0 (http://www.broad.mit.edu/mpg/sweep).

Values of the integrated haplotype score (iHS) statistic (Voight et al. 2006) for the HapMap Phase I data (http://www.hapmap.org/) and their *P* values were obtained using the online tool Haplotter (http://hg-wen.uchicago.edu/selection/haplotter.htm).

Results

To characterize the patterns of variation across the proximal *SERPIN* subcluster at 14q32.1, we surveyed 7 DNA fragments, spanning a total of 18.7 kb and covering the coding and adjacent noncoding segments depicted in figure 1. All segments were sequenced in each individual in the samples from the sub-Saharan African population of São Tomé and from the European population of Portugal, in which the common polymorphisms of *SERPINA1* had been previously typed (Seixas et al. 2001).

We also sequenced the orthologous regions in 1 chimpanzee, 1 gorilla, and 1 orangutan, to infer the ancestral states for each human polymorphism. Alignment of the human (AL117259) and chimpanzee (NW_115886) genomic reference sequences showed that an ~7.5-kb region orthologous to positions 20714–28182 in the human sequence and spanning the entire *SERPINA2* gene was absent in the chimpanzee. To evaluate whether the missing sequence represented a genuine deletion, we performed PCR sequencing in our chimpanzee sample using primers specific to regions flanking the putative deletion (AL117259 19321–19342 and 29646–29668). The results indicated that both chromosomes analyzed harbor the ~7.5-kb deletion encompassing the whole *SERPINA2* gene, suggesting that this gene deletion is not rare and is, perhaps, fixed in chimpanzee. In contrast, gorilla and orangutan sequences showed an intact *SERPINA2* gene.

Sequence Variation and Polymorphism Levels

We found a total of 130 polymorphic sites, including 14 nonsynonymous, 13 synonymous, and 103 noncoding mutations. In *SERPINA1*, we observed 5 amino acid replacements resulting in the common protein variants previously described (R101H, A213V, E264V, E342L, and E376D). In *SERPINA6*, we found a previously identified A224S polymorphism (Smith et al. 1992; Torpy et al. 2004). In *SERPINA2*, we confirmed that a 2,024-bp deletion spanning from intron III to exon V (fig. 1) is a common polymorphism, occurring in 23 and 7 chromosomes in the São Tomé and Portuguese samples, respectively. Within the deletion-carrying chromosomes, we additionally identified 1 mutation in the start codon (AT$\underline{\text{G}}$ → AT$\underline{\text{A}}$) and 1 amino acid replacement variant (Q102L). In the chromosomes not carrying the deletion, we identified 2 frameshift mutations

Table 1
Summary Statistics of Population Variation

		SERPINA1						*SERPINA2*[a]						*SERPINA6*					
	N[b]	L[c]	S[d]	θ_W[e]	π[f]	D[g]	ρ_{H01}[h]	L[c]	S[d]	θ_W[e]	π[f]	D[g]	ρ_{H01}[h]	L[c]	S[d]	θ_W[e]	π[f]	D[g]	ρ_{H01}[h]
São Tomé	40	7,587	44	13.63	11.35	−0.59	15.01	3,098	20	15.18	14.26	−0.20	3.81	6,273	42	15.74	17.58	0.41	7.61
Portugal	40	7,587	29	8.99	12.22	1.24	8.91	3,098	15	11.38	13.81	0.68	4.51	6,273	35	13.12	18.73	1.49	3.41

[a] Variation in segments spanning the 2-kb deletion were omitted from analysis.

[b] N = number of chromosomes.

[c] L = total number of sites surveyed.

[d] S = number of segregating sites.

[e] Watterson's estimator of θ ($4N_e\mu$) (Watterson 1975) per basepair ($\times 10^{-4}$).

[f] Nucleotide diversity per basepair ($\times 10^{-4}$).

[g] Tajima's D statistic (Tajima 1989).

[h] Hudson's estimator of ρ ($4N_er$) per basepair ($\times 10^{-4}$), based on a conversion-to-crossover ratio of 2 and a mean conversion tract length of 500 bp (Frisse et al. 2001; Hudson 2001).

leading to premature stop codons (L108fs and L277fs) and 4 amino acid replacement variants (I280T, L308P, E330K, and P387L). Except for I280T and E330K, these amino acid replacements are likely to alter the protein structure based on the computational predictions of Polyphen (Ramensky et al. 2002). In the sample from São Tomé, the overall frequency of chromosomes bearing mutations likely to affect *SERPINA2* function is 85% (34/40) (i.e., 57.5% deletion, 10% frameshift mutations, and 17.5% mutations predicted to affect protein structure). In the Portuguese sample, the frequency is 67.5% (27/40) (i.e., 17.5% deletion, 7.5% frameshift mutations, and 42.5% mutations predicted to affect protein structure).

Summary statistics of the polymorphism and sequence divergence data are shown in table 1. Variation within the deleted fragment of *SERPINA2* was omitted from this analysis. Polymorphism levels as summarized by nucleotide diversity (π), which is based on average number of differences between sequences, and by the estimator of the population mutation rate parameter θ_W (Watterson 1975), which is based on the number of polymorphic sites and sample size, are slightly higher than the genome-wide average in humans (Crawford et al. 2005; Stajich and Hahn 2005). The São Tomé sample shows the highest number of polymorphic sites in all surveyed regions. Estimates of the population recombination rate parameter ($4N_er$) based on the composite likelihood estimator ρ_{H01} (Frisse et al. 2001; Hudson 2001) are higher than genome-wide average values (4×10^{-4} and 2×10^{-4} for African and European–Americans, respectively) (Serre et al. 2005) and fall within the range estimated for subtelomeric regions (Serre et al. 2005). These results are consistent with the high recombination rates (2 cM/Mb) reported for the interval spanning the *SERPIN* cluster (Kong et al. 2002). The lowest and highest levels of linkage disequilibrium (LD), as summarized by ρ_{H01}, were found in the *SERPINA1* and *SERPINA2* regions, respectively. *SERPINA2* was the only region in which LD levels in São Tomé were not lower than in the Portuguese sample (table 1). The Tajima's D statistic, which summarizes information about the spectrum of allele frequencies, is expected to be approximately 0 under the neutral equilibrium model (Tajima 1989). A negative value indicates an excess of rare variants, which may result from a selective sweep, whereas a positive value indicates an excess of intermediate frequency variants, which may reflect the action of balancing selection. Studies of sequence variation in humans have shown that populations of African ancestry tend to have a slight excess of rare variants, whereas non-African populations show an excess of intermediate frequency variants (Wall and Przeworski 2000; Frisse et al. 2001; Akey et al. 2004; Stajich and Hahn 2005; Voight et al. 2005). All Tajima's D values obtained for the *SERPIN* subcluster do not depart significantly from the expectations of the neutral equilibrium model and show the same trends observed in population samples from similar ethnic groups (Wall and Przeworski 2000; Frisse et al. 2001; Akey et al. 2004; Stajich and Hahn 2005; Voight et al. 2005).

Haplotype Diversity and Tests for the Signature of Natural Selection

A visual representation of the sequence data in the form of inferred haplotypes is shown in figure 2. In the São Tomé sample, the *SERPINA2* region harbors a common haplotype defined by the 2-kb deletion and by derived alleles at 3 additional sites (25095, 26806, and 26911). Sites 23876 (corresponding to the start codon AT$\underline{\text{G}}$ → AT$\underline{\text{A}}$ mutation) and 25546 were also found to bear derived alleles that are strongly associated with the 2-kb deletion ($|D'| = 1$; $r^2 = 0.82$ and $|D'| = 0.89$; $r^2 = 0.71$, respectively).

To evaluate whether this haplotype structure could result from the action of positive natural selection, we calculated the EHH statistic proposed by Sabeti et al. (2002). Specifically, we measured the decay of LD around core haplotypes defined by sites 26806 and 26911, which are surrogate markers for the 2-kb deletion polymorphism at *SERPINA2*. When EHH was plotted against distance in the São Tomé sample, we found that, in spite of their higher frequency (0.575 vs. 0.425), chromosomes with the 26806–26911 G-A haplotypes, bearing the 2-kb deletion, have greater EHH than T-G chromosomes bearing the nondeleted allele (fig. 3*A*). In the Portuguese sample, we used 3 additional polymorphic sites (25881, 26354, and 26461) to define a set of core haplotypes that occur at frequencies closer to the deletion haplotype (fig. 3*B*). In this case, EHH for the 2-kb deletion haplotype (CGTGA) still appears to decay relatively slowly compared with the EHH for other intermediate frequency haplotypes (CATTG and CGATG).

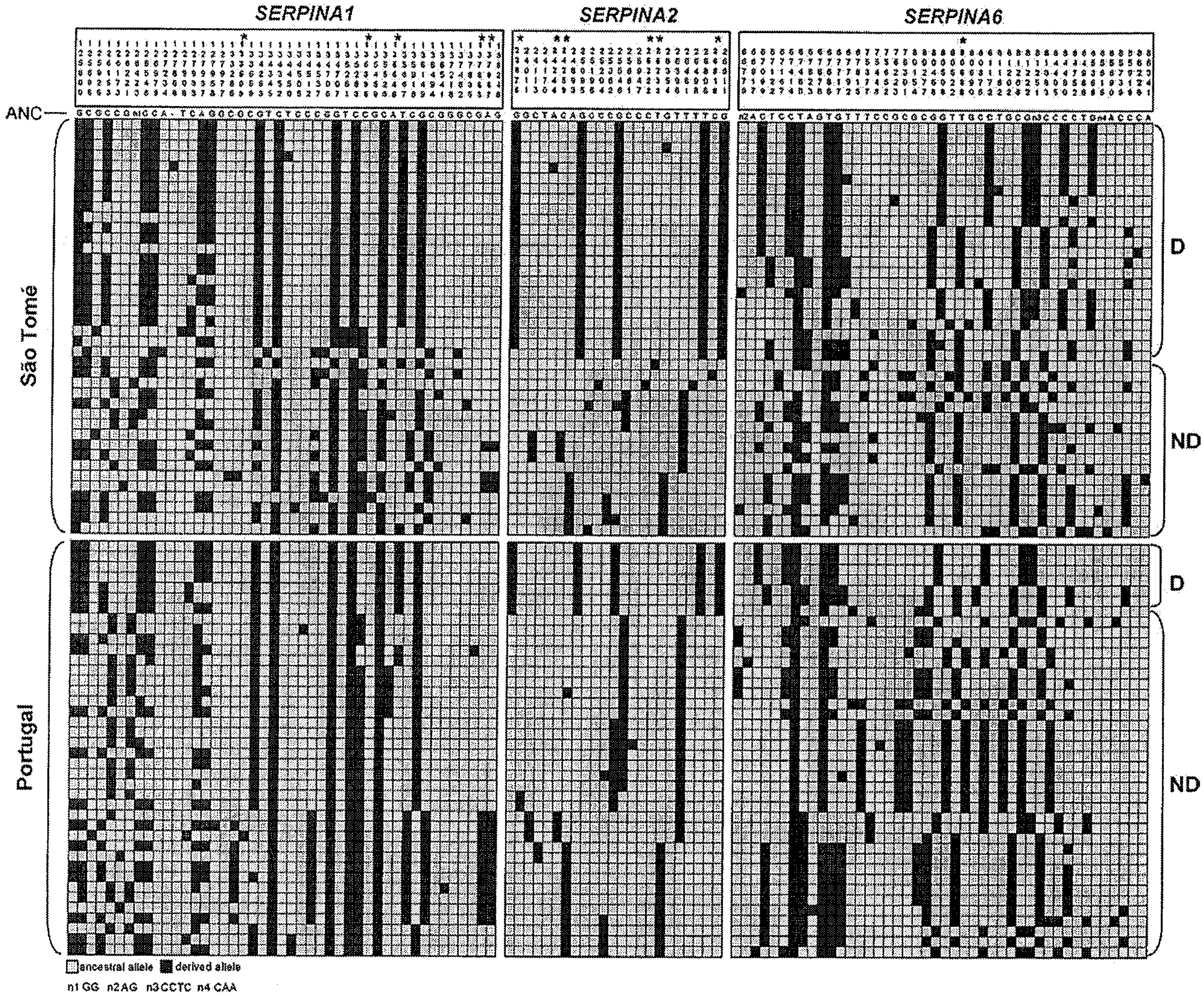


FIG. 2.—Haplotypes inferred by PHASE for *SERPINA1*, *SERPINA2*, and *SERPINA6*. The nonhuman primate sequences were used to infer the ancestral state at each site. Numbers below the gene names indicate the position of each polymorphic site relative to the reference sequence for each gene (GenBank accession numbers AL132708 for *SERPINA1* and AL117259 for *SERPINA2* and *SERPINA6*). The segment spanned by the 2-kb deletion was omitted. Nonsynonymous sites are marked by asterisk. D and ND indicate the chromosomes bearing the *SERPINA2* deletion and nondeletion alleles, respectively.

To test the null hypothesis of evolutionary neutrality for the *SERPINA2* haplotypes, we used the haplotype test described by Hudson et al. (1994) and the LRH test proposed by Sabeti et al. (2002).

To evaluate if the haplotype class defined by the 2-kb deletion and by derived alleles at sites 25095, 26806, and 26911 contained fewer segregating sites than expected under neutrality, given its frequency the haplotype test of Hudson et al. (1994) was applied to windows of different sizes defined by the concatenation of the different surveyed fragments (fig. 1). Using the haplotypes inferred by PHASE as the best reconstruction, we obtained significant tests in the São Tomé sample, for the 8-kb region resulting from concatenation of fragments F2 and F3 (P = 0.0036; fig. 1) and the 10.5-kb region including fragments F1–F3 (P = 0.0275). In the Portuguese sample, we tested the same concatenated fragments, but no test was found to be significant (P values of 0.74 and 0.98, respectively). Haplotype tests of fragments centered on *SERPINA6* were not significant in either population sample (P values ranging from 0.14 to 0.99). All tests were based on simulations of the standard neutral model. This model was shown to provide a good fit to sequence variation data for populations of African ancestry, including the admixed populations of African-Americans (Adams and Hudson 2004; Voight et al. 2005). Conversely, non-African data do not fit the standard neutral model and were shown to be compatible with bottleneck models (Adams and Hudson 2004; Voight et al. 2005). Because the evolutionary variance is greater under bottleneck models, it is highly likely that the tests of the Portuguese data would remain not significant even if the simulations were based on such models.

To determine if test results in the São Tomé sample were robust to misspecification of haplotype assignment, we generated 100 haplotype samples drawn at random from the posterior distribution estimated by PHASE. After rerunning the test on each sample, we found that for the 8-kb concatenated fragment including F2–F3, all samples had

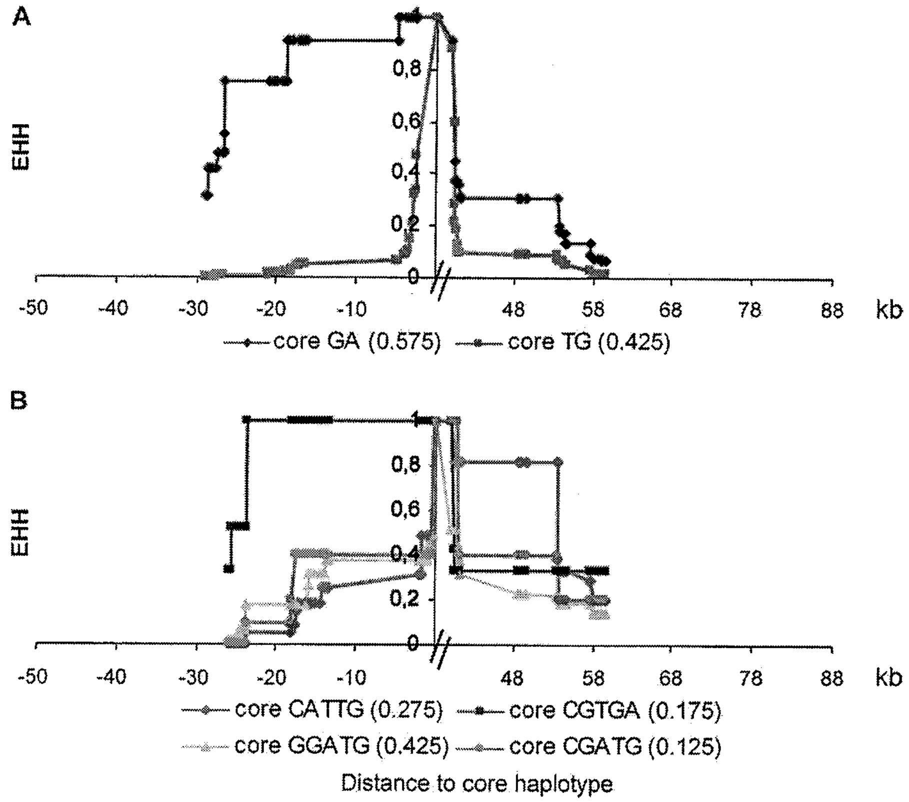


FIG. 3.—Plots of EHH breakdown over distance from core haplotypes defined by SNPs in *SERPINA2*. (*A*) Comparison between São-Tomean chromosomes bearing the 2-kb deletion (GA) and nondeletion (TG) alleles at *SERPINA2*. (*B*) Comparison between Portuguese chromosomes bearing the 2-kb deletion (CGTGA) and nondeletion (CATTG, GGATG, and CGATG) alleles. The unsurveyed segment between *SERPINA2* and *SERPINA6* was omitted. Frequencies of core haplotypes are shown in parenthesis.

a *P* value lower than 0.01, and for the 10.5-kb region including fragments F1–F3, 96% of samples had *P* values lower than 0.05. Taken together, these results suggest that the São Tomé sample harbors a signature of natural selection on a haplotype class that encompasses a region up to 28 kb in length, including *SERPINA1* and *SERPINA2*.

To determine whether the EHH for the 26806–26911 G-A core haplotype in the African sample was unusual, we compared the haplotype frequency with the relative EHH at the 2 largest distances where non–G-A haplotypes had nonzero values of EHH (−25 kb distal and 60 kb proximal; fig. 3*A*). The deviations from simulated null distributions were significant for all 4 tested demographic models (*P* values for −25 kb distal are constant-sized population, $P < 0.000007$; expansion, $P < 0.000008$; bottleneck, $P < 0.0000005$; and population structure, $P < 0.00001$; and those for 60 kb proximal are constant-sized population, $P < 0.02$; expansion $P < 0.004$; bottleneck $P < 0.01$; and population structure $P < 0.003$).

Simulation-based tests allow testing the hypothesis of evolutionary neutrality based on specified demographic scenarios. An alternative, empirical approach consists in comparing the haplotype structure observed at a candidate locus with the pattern observed over the entire genome. This approach was recently implemented for the HapMap Phase I data (http://www.hapmap.org/) by calculating an iHS for each single nucleotide polymorphism (SNP) in each of the 3 HapMap population samples; the absolute value of iHS measures the strength of the evidence for selection acting on a SNP or one tightly linked to it (Voight et al. 2006). In the Yoruba, the |iHS| for SNP rs6647 (site 135395), which is the best surrogate marker for the *SERPINA2* deletion in HapMap Phase I data, is 2.611; this indicates that, consistent with the resequencing data, SNP rs6647 is associated with a selection signal in the top 5% of the entire genome for this population. However, because this SNP is only moderately correlated with the deletion allele ($r^2 = 0.22$), it may not be ideal for assessing the evidence for selection acting on the deletion. Therefore, we used the iHS statistic to ask whether the SERPIN genes evolved by positive selection. In this analysis, the signal is proportional to the number of SNPs with $|iHS| > 2$ in a window of 50 SNPs centered on the gene (Voight et al. 2006). Based on this approach, the empirical *P* values for the *SERPIN* genes in the Yoruba sample also appear to be extreme in the genome-wide distribution (*SERPINA1* $P = 0.053$, *SERPINA2* $P = 0.053$, and *SERPINA6* $P = 0.037$), further suggesting a recent selective process. It should be noted that the iHS results point to a stronger signal in *SERPINA6* compared with *SERPINA1* and *SERPINA2*, in contrast with what we observed in our resequencing study. However, the HapMap Phase I data only included a few SNPs within the proximal *SERPIN* subcluster (8 SNPs within *SERPINA1*; 2 SNPs within *SERPINA2*; and 9 SNPs within *SERPINA6*) and, as a consequence, the SNP windows analyzed for each gene largely overlap. Hence, it

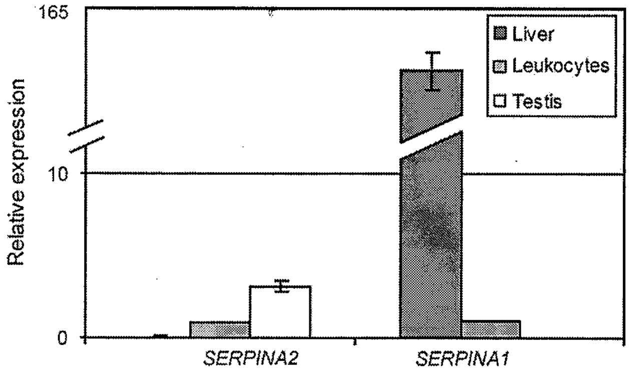


FIG. 4.—Relative mRNA levels of SERPINA2 and SERPINA1 in liver, leukocytes, and testis, as estimated by quantitative RT PCR.

is not possible to determine which gene contributes to the selection signal based on the HapMap data.

In the European (Centre d'Etude du Polymorphisme Humain—Utah residents with ancestry from northern and western Europe) and Asian (Han Chinese from Beijing and Japanese from Tokyo) population samples, no signal of selection was detected based on the |iHS| approach, corroborating the idea that the signature of selection is specific to African populations.

Expression Studies SERPINA2 and Evaluation of Protein Structure

Although the haplotype structure suggests a role for positive selection in shaping variation at the *SERPIN* subcluster, the target of this selection was not immediately obvious. Site 135395, corresponding to the A213V amino acid replacement at *SERPINA1* could be regarded as a potential target of selection. However, this seems unlikely because: 1) the allele associated with the selected haplotype is the ancestral one (A213) (see fig. 2) and 2) the A213 allele is only loosely linked to the selected haplotype ($r^2 = 0.22$). Moreover, clinical studies did not report any phenotypic association with the A213 allele that would suggest a selective advantage (Nukiwa et al. 1987; Gaillard et al. 1994). With regard to *SERPINA6*, a single nonsynonymous variant was observed (A224S) and this variant is not strongly associated with the selected haplotype ($r^2 = 0.11$).

Hence, the remaining possible target of selection is the deletion of the *SERPINA2* gene. However, if the nondeleted form of *SERPINA2* is not a functional gene, as previously proposed, the deletion would be unlikely to have phenotypic and fitness effects. We tested for *SERPINA2* expression in liver, testis, and leukocytes and used RT PCR to assess the expression of *SERPINA2* relative to *SERPINA1*. The *SERPINA2* gene is highly expressed in the testis at 3-fold and 60-fold higher levels relative to leukocytes and liver, respectively (fig. 4). On the contrary, and as previously reported, the highest expression levels of *SERPINA1* were observed in the liver, the major SERPINA1 producer (Cox 1995; Nukiwa et al. 1996), at 160-fold higher levels compared with leukocytes. In testis, multiple attempts did not show evidence for *SERPINA1* expression. Interestingly, leukocytes appeared to have higher expression levels of *SERPINA2* (28 ct) than *SERPINA1* (34 ct). These results demonstrate that the nondeleted form of *SERPINA2* is expressed and its expression is different across the tissues tested. In addition, *SERPINA1* and *SERPINA2* are differentially expressed relative to each other.

To further explore the possibility of a functional role for SERPINA2, a 3-dimensional protein model (http://swissmodel.expasy.org/workspace/) for the nondeleted form was built on the scaffold of the available crystal structure of the highly homologous SERPINA1 protein (75–81% homology). According to this model (fig. 5), SERPINA2 preserves the typical structure of the SERPIN reactive center loop, which is compatible with a protease inhibitory activity (fig. 5*B*). However, the sequences flanking the reactive site appear to have diverged considerably (fig. 5*C*), and the putative reactive site (P1–P1′) of SERPINA2 harbors a tryptophan–serine sequence instead of a methionine–serine motif, as previously noted (Bao et al. 1988). The tryptophan–serine motif was also found in the orthologous sequences of our gorilla and orangutan samples as well as in the rhesus monkey (UCSC Genome Browser,—http://genome.ucsc.edu/—chr7: 157590203–157595117) and in a rat SERPINA2 putative protein (XP_2345123). Moreover, other members of the SERPIN family share a reactive site similar to that of SERPINA2 (P1-hydrophobic and P1′-polar amino acids); they include SERPINA10 (Q5RDA8), SERPINA4 (P29622), mouse α-1-antichymotrypsin (P01011), and chicken ovalbumin (P01012). Hence, it appears that the nondeleted SERPINA2 protein has retained important features of a functional active site.

Discussion

We have performed an extensive survey of sequence variation of the proximal 14q32.1 *SERPIN* subcluster, including the 2 functional genes, *SERPINA1* and *SERPINA6*, and the previously proposed pseudogene, *SERPINA2*. In the São Tomé sample, we show that a haplotype defined by a partial deletion of *SERPINA2* is associated with too little variation, given its frequency, relative to neutral expectations. We further show that, in the absence of the deletion allele, the *SERPINA2* gene may be active and differentially expressed in liver, testis, and leukocytes. Moreover, its expression pattern is distinct from that of the highly homologous *SERPINA1* gene. Finally, we show that a 7.5-kb deletion removed the *SERPINA2* ortholog in the chimpanzee genome, but not in the other great apes. Hence, we propose that the loss of *SERPINA2* was advantageous in recent primate evolution.

We assessed the signature of positive natural selection in 2 ways. The first one was based on the application of the haplotype test of Hudson et al. (1994) and the LRH test (Sabeti et al. 2002) to our full resequencing data and relied on the assumption of a specified set of demographic models. By this approach, we found a homogeneous haplotype class, particularly frequent in the African sample that is defined by the 2-kb deletion, an $AT\underline{G} \rightarrow AT\underline{A}$ mutation in the start codon and 3 additional mutations in near-perfect LD with each other. This finding suggests that a haplotype class was quickly driven to high frequency by natural selection acting on an advantageous variant. Although the São Tomé sample is geographically defined, the population is likely to

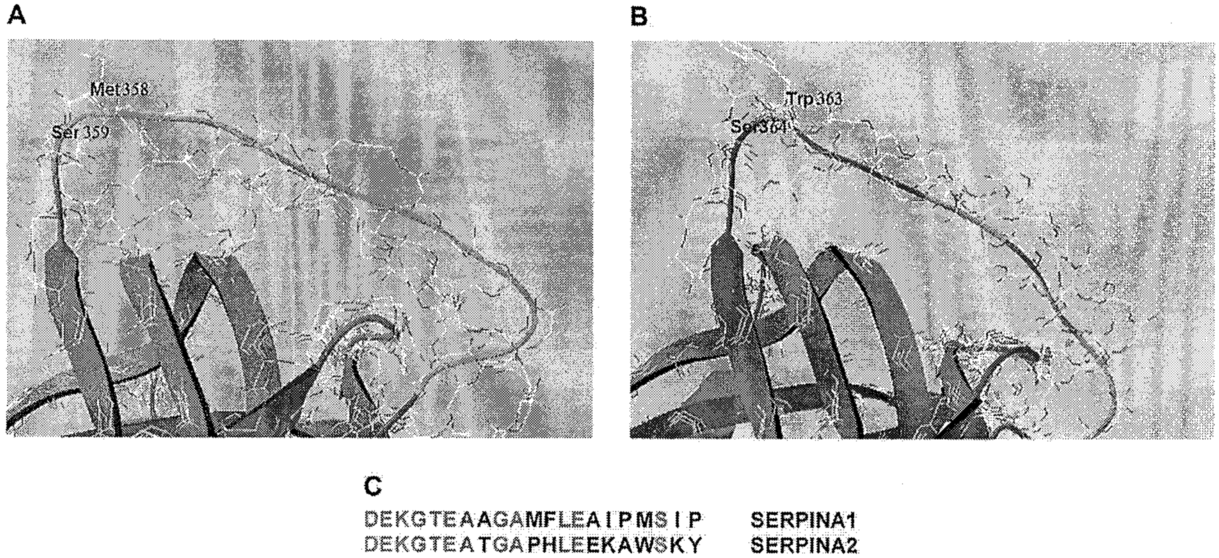

A
Met358
Ser 359
B
Trp363
Ser364
C
DEKGTEAAGAMFLEAIPMSIP SERPINA1
DEKGTEATGAPHLEEKAWSKY SERPINA2

FIG. 5.—Three-dimensional theoretical model of the SERPIN reactive center loop (RCL), as obtained from SWISS-MODEL Automated Protein Modelling Server. (*A*) SERPINA1 RCL (D341–P361). (*B*) SERPINA2 RCL (D346–Y366). (*C*) RCL sequence homology between the SERPINA1 and SERPINA2 (conserved amino acids are shown in red). Images were obtained using the Swiss PDB viewer 3.7 (http://www.expasy.ch/spdbv/).

be a mixture of people from different parts of sub-Saharan Africa (Tomas et al. 2002). Hence, we cannot definitively rule out the possibility that the observed departure in the haplotype test is due to a violation of the demographic assumptions of the standard neutral model rather than positive natural selection. However, previous multilocus surveys of sequence variation in samples of sub-Saharan African ancestry, including the highly admixed African-Americans, did not detect significant departures from the standard neutral model (Adams and Hudson 2004; Akey et al. 2004; Stajich and Hahn 2005; Voight et al. 2005). Moreover, the significance of the results remained unchanged under alternative demographic models, including population subdivision. This suggests that simulation-based tests for the São Tomé data may be relatively robust. The second approach we used to detect the signature of natural selection, however, does not suffer from this potential problem and offers additional advantages. In this approach, we used the iHS statistic to investigate the signature of natural selection in the HapMap Phase I data; the iHS has high power to detect a sweep of a variant at intermediate frequency such as the *SERPINA2* deletion. Because the HapMap samples include a well-defined African population, that is, the Yoruba from Ibadan (Nigeria), admixture is a less serious problem. In addition, the analysis based on the iHS does not rely on the assumptions of a null neutral model in that the data at the test locus, in this case *SERPINA2*, are compared with the empirical genome-wide distribution in the same population sample. This approach effectively circumvents the need to specify a model of population history that is unknown. On the other hand, a possible drawback of genome scans for selection, such as those performed by using the iHS statistics, is that they may have limited power because only the strongest selection signals can be detected and the false negative rate may be high (Teshima et al. 2006). Despite this potential limitation, the genes in the proximal *SERPIN* subcluster exhibit signals of selection that are in the top 3.7–5.3% of the distribution for the HapMap Phase I data in the Yoruba. We also performed an SNP-based analysis of the iHS statistic by using a surrogate SNP for the deletion allele (rs6647); the iHS for this SNP in the Yoruba is 2.611, corresponding to an empirical *P* value of 0.05–0.01. Thus, both approaches and both data sets converge on supporting the notion that the haplotype structure at the proximal subcluster is unusual and likely to be due to the action of natural selection acting on an advantageous variant. This conclusion is consistent with our previous analysis of a microsatellite located between *SERPINA1* and *SERPINA2*, which exhibited lower levels of diversity in a sample from São Tomé compared with 2 European samples (from Portugal and Basque Country) as well as a unimodal allele frequency distribution within the A213 (rs6647) chromosomes from São Tomé (Seixas et al. 2001). Consistent with the findings reported here, the patterns of microsatellite variation were proposed to be the result of positive natural selection on a unknown advantageous variant tightly linked to *SERPINA1* (Seixas et al. 2001).

Full resequencing of confined genomic regions and genome-wide SNP typing offer complementary opportunities for detecting the signature of natural selection. Although the latter allows the assessment of haplotype homozygosity over large distances and, hence, may have greater power to detect a signature of selection for less common variants, it provides limited information on the underlying patterns of variation and the full array of variants. On the other hand, thanks to the availability of large-scale data sets, SNP-typing data allow the implementation of empirical approaches that do not rely on assumptions about the unknown underlying demographic model. In the case of the proximal *SERPIN* subcluster, there is generally good agreement between these 2 types of data for the population samples of African ancestry. In the HapMap data, the *SERPINA6* gene has a slightly stronger signal of selection compared with the other 2 genes in the subcluster ($P = 0.037$ vs. $P = 0.053$ at *SERPINA6* vs. *SERPINA2*, respectively) corresponding to a larger fraction of SNPs with |iHS| > 2. This gene-based analyses, as described and implemented in Voight et al. (2006), use information from a window of at least 50 SNPs centered on each gene; because of the close proximity of *SERPINA6* to *SERPINA2* and the sparseness of the HapMap Phase I data, many of the SNPs in the windows for these 2 genes overlap, making it difficult to distinguish the location of the selection signal. In our resequencing data, the signature of selection seems to be stronger for a region centered on *SERPINA2*. Indeed, haplotype tests centered on *SERPINA6* did not yield significant results. Because of the higher density of the resequencing data and the presence of a common variant with a clear effect on gene function, that is, the 2-kb deletion in *SERPINA2*, we hypothesize that *SERPINA2* was the target of selection. The other damaging mutations observed in both samples (i.e., 2 frameshift mutations and 2 mutations predicted to affect function) may also be advantageous. The 2 amino acid replacements are expected to have less drastic effects on function compared with the deletion; therefore, they may not be as strongly advantageous as the 2-kb deletion. The 2 frameshift mutations are relatively rare (7.5% and a singleton) and may be more recent than the 2-kb deletion. The finding of 2 independent rearrangements, namely the 2-kb deletion in humans and the 7.5-kb deletion in chimpanzees, in the 6 Myr elapsed since the divergence of these 2 species, corroborates the idea that loss of *SERPINA2* is an ongoing process that was associated with a selective advantage during recent primate evolution.

This hypothesis is consistent with the high mRNA levels and the tissue-specific expression profiles observed for the nondeleted *SERPINA2* gene. In addition, the analysis of structural models shows that the nondeleted *SERPINA2* gene not only is expressed, but may also code for a functional SERPIN. Interestingly, although both SERPINA1 and SERPINA2 have the prototypical SERPIN configuration, they diverged at the reactive center, which may have led to different substrate affinities and SERPIN activities.

It was previously proposed that, in some cases, pseudogenization could confer a selective advantage. For example, a variant with a premature stop codon in the *Caspase12* gene, which codes for a cystein protease, reached near-fixation frequency probably because it confers increased resistance to severe sepsis (Wang et al. 2006; Xue et al. 2006). Likewise, the G protein–coupled receptor 33 (*GPR33*) and CMP-N-acetylneuraminic acid hydroxylase (*CMAH*) genes also carry inactivating alleles at high frequencies (a premature stop codon in *GPR33* and a 92-bp deletion removing *CMAH* exon 6) (Rompler et al. 2005; Hayakawa et al. 2006). These inactivating mutations were estimated to have occurred 1 MYA and 3 MYA, respectively. Interestingly, the loss of *GPR33* function was proposed to have occurred multiple times during primate and rodent evolution probably as a result of a shared selective pressure (Rompler et al. 2005). Hence, our findings on the loss of *SERPINA2*, resulting from deletions and other inactivating mutations, add to a growing body of evidence supporting the idea that pseudogenization was advantageous in recent human evolution. Several recent genome-wide surveys have shown that polymorphic deletions are a pervasive feature of human variation (Sebat et al. 2004; Sharp et al. 2005; Tuzun et al. 2005; Conrad et al. 2006; Feuk et al. 2006; Hinds et al. 2006; McCarroll et al. 2006). Despite their abundance, only a few examples of such variants were shown to affect risk for common disease phenotypes or play an important role in human adaptations. In this sense, the polymorphic deletion of *SERPINA2* may provide another example of the evolutionary potential of such variation. More generally, if loss of function is an important mechanism contributing to species differences, as posited by the less is more hypothesis (Olson 1999), deletions may be an important source of advantageous variation. Interestingly, in the case of *SERPINA2*, there seems to be an adaptive convergence between humans and chimpanzees toward loss of function.

Although additional functional studies are needed to determine whether the *SERPINA2* deletion has phenotypic effects, the high expression levels found in the testis raise the possibility that SERPINA2 plays a role in reproduction. Proteins implicated in reproduction have already been documented in several taxa as preferred targets for adaptive evolution, driven by sperm competition, sexual conflict, or pathogen interactions (Clark and Swanson 2005). It is well known that the orchestrated interaction between proteases and protease inhibitors plays a crucial role in sperm modifications and in tissue integrity, from spermatogenesis to fertilization. In rodents and primates, SERPINs have the ability to inhibit proteases found in the reproductive tract, including semen proteases, such as urokinase (uPA) and/or kallikrein 3 (KLK3 or prostate-specific antigen), which liquefy sperm coagulum leading to the release of sperm. Consistent with our observation at *SERPINA2*, comparative genomic analysis of proteases and protease inhibitors in rodents (mouse and rat) and primates (human and chimpanzee) have shown a marked change in gene content, namely in *kallikreins* and *SERPIN* cluster genes, with lineage-specific patterns of gene inactivation that were proposed to reflect differences in reproductive biology (Puente and Lopez-Otin 2004; Puente, Gutierrez-Fernandez, et al. 2005; Puente, Sanchez, et al. 2005).

Alternatively, the signature of selection on the haplotype carrying the *SERPINA2* deletion may result from a selective pressure mediated through host–pathogen interactions. In fact, proteases and their inhibitors are known to have important roles in host defense against pathogens, within and outside the reproductive tract. For example, prolactin-induced protein is a protease from the seminal fluid that is thought to protect sperm from infection by binding bacteria and suppressing T-cell apoptosis (Schenkels et al. 1997; Gaubin et al. 1999). On the other hand, SERPINs are proposed to play an important role in antagonizing the pathogen invasion process (Hill and Hastie 1987; Goodwin et al. 1996).

For example, SERPINA1 was shown to interfere with Schistosoma, Cryptosporidium, and HIV infections (Asch and Dresden 1977; Forney et al. 1996, 1997; Shapiro et al. 2001; Freudenstein-Dan et al. 2003; Hayes and Gardiner-Garden 2003). The expression of *SERPINA2* in leukocytes is consistent with a role in the response to pathogens. Infectious diseases have constituted a major selective pressure in human populations, especially in concomitance with the environmental changes linked to the onset of agriculture.

For example, the onset of malaria in Africa has resulted in strong selective pressures acting on variation at the Duffy blood group, β-globin, and glucose-6-phosphate dehydrogenase loci (Hamblin and Di Rienzo 2000; Tishkoff et al. 2001; Currat et al. 2002; Seixas et al. 2002). Interestingly, a crude estimate of the time to the most recent common ancestor of the deletion allele based on the decay of haplotype homozygosity by historical recombination (Voight et al. 2006) suggests a recent selective event, that is, 10,000–24,000 years ago, consistent with the history of selective pressures on host–pathogen interactions.

We used a surrogate marker for the deletion, that is, rs1956172 ($r^2 = 1$ and $r^2 = 0.82$ in Portugal and São Tomé, respectively), to infer the deletion frequencies in the HapMap phase II (http://www.hapmap.org/) and Perlegen data (http://genome.perlegen.com/) (Hinds et al. 2005). The inferred frequencies are highest in samples of African ancestry (0.52 in Yoruba and 0.41 in African-Americans), intermediate in Europeans (0.23 in Central Europeans from Utah and 0.23 in European-Americans), and lowest in those of Asian ancestry (0.01 in Chinese, 0 in Japanese, and 0.06 in individuals of Han Chinese ancestry). Based on this geographic distribution of allele frequencies and the restriction of the selection signature to the African samples, it may be speculated that an adaptive pressure driven by host–pathogen interactions is more likely than a selective advantage due to an effect on fertility. Additional data on the phenotypic consequences of the *SERPINA2* deletion are necessary to further test the hypothesis of a selective advantage for SERPINA2 loss in human populations.

Acknowledgments

The authors wish to thank the people and the Ministry of Health of the Democratic Republic of São Tomé e Príncipe, Maria de Jesus Trovoada, and Licínio Manco for making São Tomé blood samples available for DNA extraction, Alberto Barros and Mário Sousa for having provided biopsies from human testis, and Molly Przeworski for comments on the manuscript. S.S. is the recipient of a postdoctoral fellowship from the Fundação para a Ciência e a Tecnologia, Portugal (FCTBPD/12532/2003). This research was supported by the Fundação para a Ciência e a Tecnologia (POCTI/42510/ANT/2001 and POCI/BIA-BDE/56654/2004 grants to J.R.) and by National Institutes of Health grant DK56670 to A.D.

Literature Cited

Adams AM, Hudson RR. 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics. 168:1699–1712.

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. 2:e286.

Asch HL, Dresden MH. 1977. Schistosoma mansoni: inhibition of cercarial "penetration" proteases by components of mammalian blood. Comp Biochem Physiol B. 58:89–95.

Atchley WR, Lokot T, Wollenberg K, Dress A, Ragg H. 2001. Phylogenetic analyses of amino acid variation in the serpin proteins. Mol Biol Evol. 18:1502–1511.

Bao JJ, Reed-Fourquet L, Sifers RN, Kidd VJ, Woo SL. 1988. Molecular structure and sequence homology of a gene related to alpha 1-antitrypsin in the human genome. Genomics. 2: 165–173.

Carrell RW, Lomas DA. 2002. Alpha1-antitrypsin deficiency—a model for conformational diseases. N Engl J Med. 346:45–53.

Clark NL, Swanson WJ. 2005. Pervasive adaptive evolution in primate seminal proteins. PLoS Genet. 1:e35.

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. 2006. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet. 38:75–81.

Cox DW. 1995. α1-antitrypsin deficiency. In: Sriver CR, Beaudet AL, Sly WL, Valle D, editors. The metabolic and molecular bases of inherited disease. New York: McGraw-Hill. p. 4125–4158.

Crawford DC, Akey DT, Nickerson DA. 2005. The patterns of natural variation in human genes. Annu Rev Genomics Hum Genet. 6:287–312.

Creighton TE, Darby NJ. 1989. Functional evolutionary divergence of proteolytic enzymes and their inhibitors. Trends Biochem Sci. 14:319–324.

Currat M, Trabuchet G, Rees D, Perrin P, Harding RM, Clegg JB, Langaney A, Excoffier L. 2002. Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta(S) Senegal mutation. Am J Hum Genet. 70:207–223.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet. 7:85–97.

Forney JR, Yang S, Healey MC. 1996. Protease activity associated with excystation of Cryptosporidium parvum oocysts. J Parasitol. 82:889–892.

Forney JR, Yang S, Healey MC. 1997. Synergistic anticryptosporidial potential of the combination alpha-1-antitrypsin and paromomycin. Antimicrob Agents Chemother. 41:2006–2008.

Freudenstein-Dan A, Gold D, Fishelson Z. 2003. Killing of schistosomes by elastase and hydrogen peroxide: implications for leukocyte-mediated schistosome killing. J Parasitol. 89: 1129–1135.

Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet. 69: 831–843.

Gaillard MC, Zwi S, Nogueira CM, Ludewick H, Feldman C, Frankel A, Tsilimigras C, Kilroe-Smith TA. 1994. Ethnic differences in the occurrence of the M1(ala213) haplotype of alpha-1-antitrypsin in asthmatic and non-asthmatic black and white South Africans. Clin Genet. 45:122–127.

Gaubin M, Autiero M, Basmaciogullari S, Metivier D, Misëhal Z, Culerrier R, Oudin A, Guardiola J, Piatier-Tonneau D. 1999. Potent inhibition of CD4/TCR-mediated T cell apoptosis by a CD4-binding glycoprotein secreted from breast tumor and seminal vesicle cells. J Immunol. 162:2631–2638.

Goodwin RL, Baumann H, Berger FG. 1996. Patterns of divergence during evolution of alpha 1-proteinase inhibitors in mammals. Mol Biol Evol. 13:346–358.

Hamblin MT, Di Rienzo A. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet. 66:1669–1679.

Hayakawa T, Aki I, Varki A, Satta Y, Takahata N. 2006. Fixation of the human-specific CMP-N-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. Genetics. 172:1139–1146.

Hayes VM, Gardiner-Garden M. 2003. Are polymorphic markers within the alpha-1-antitrypsin gene associated with risk of human immunodeficiency virus disease? J Infect Dis. 188: 1205–1208.

Hill RE, Hastie ND. 1987. Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature. 326:96–99.

Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet. 38:82–85.

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. Science. 307:1072–1079.

Hofker MH, Nelen M, Klasen EC, Nukiwa T, Curiel D, Crystal RG, Frants RR. 1988. Cloning and characterization of an alpha 1-antitrypsin like gene 12 KB downstream of the genuine alpha 1-antitrypsin gene. Biochem Biophys Res Commun. 155: 634–642.

Hudson RR. 2001. Two-locus sampling distributions and their application. Genetics. 159:1805–1817.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18: 337–338.

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ. 1994. Evidence for positive selection in the superoxide dismutase (Sod) region of Drosophila melanogaster. Genetics. 136:1329–1340.

Irving JA, Pike RN, Lesk AM, Whisstock JC. 2000. Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. Genome Res. 10:1845–1864.

Irving JA, Steenbakkers PJ, Lesk AM, Op den Camp HJ, Pike RN, Whisstock JC. 2002. Serpins in prokaryotes. Mol Biol Evol. 19:1881–1890.

Kelsey GD, Parkar M, Povey S. 1988. The human alpha-1-antitrypsin-related sequence gene: isolation and investigation of its expression. Ann Hum Genet. 52:151–160.

Kong A, Gudbjartsson DF, Sainz J, et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. Nat Genet. 31:241–247.

Lomas DA, Carrell RW. 2002. Serpinopathies and the conformational dementias. Nat Rev Genet. 3:759–768.

Marsden MD, Fournier RE. 2005. Organization and expression of the human serpin gene cluster at 14q32.1. Front Biosci. 10:1768–1778.

McCarroll SA, Hadnott TN, Perry GH, et al. (11 co-authors). 2006. Common deletion polymorphisms in the human genome. Nat Genet. 38:86–92.

Namciu SJ, Friedman RD, Marsden MD, Sarausad LM, Jasoni CL, Fournier RE. 2004. Sequence organization and matrix attachment regions of the human serine protease inhibitor gene cluster at 14q32.1. Mamm Genome. 15:162–178.

Needham M, Stockley RA. 2004. Alpha 1-antitrypsin deficiency. 3: clinical manifestations and natural history. Thorax. 59: 441–445.

Nickerson DA, Tobe VO, Taylor SL. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res. 25:2745–2751.

Nukiwa T, Brantly M, Ogushi F, Fells G, Satoh K, Stier L, Courtney M, Crystal RG. 1987. Characterization of the M1(Ala213) type of alpha 1-antitrypsin, a newly recognized, common "normal" alpha 1-antitrypsin haplotype. Biochemistry. 26:5259–5267.

Nukiwa T, Seyama K, Kira S. 1996. The prevalence of a1AT deficiency outside the United States and Europe. In: Crystal RG, editor. Alpha-1-antitrypsin deficiency. Biology, pathogenesis, clinical manifestations, therapy. New York: Marcel Dekker Inc. p. 293–300.

Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet. 64:18–23.

Puente XS, Gutierrez-Fernandez A, Ordonez GR, Hillier LW, Lopez-Otin C. 2005. Comparative genomic analysis of human and chimpanzee proteases. Genomics. 86:638–647.

Puente XS, Lopez-Otin C. 2004. A genomic analysis of rat proteases and protease inhibitors. Genome Res. 14:609–622.

Puente XS, Sanchez LM, Gutierrez-Fernandez A, Velasco G, Lopez-Otin C. 2005. A genomic view of the complexity of mammalian proteolytic systems. Biochem Soc Trans. 33: 331–334.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 30:3894–3900.

Rompler H, Schulz A, Pitra C, Coop G, Przeworski M, Paabo S, Schoneberg T. 2005. The rise and fall of the chemoattractant receptor GPR33. J Biol Chem. 280:31068–31075.

Rozas J, Rozas R. 1997. DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput Appl Biosci. 13:307–311.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature. 419:832–837.

Schenkels LC, Walgreen-Weterings E, Oomen LC, Bolscher JG, Veerman EC, Nieuw Amerongen AV. 1997. In vivo binding of the salivary glycoprotein EP-GP (identical to GCDFP-15) to oral and non-oral bacteria detection and identification of EP-GP binding species. Biol Chem. 378:83–88.

Sebat J, Lakshmi B, Troge J, et al. (21 co-authors). 2004. Large-scale copy number polymorphism in the human genome. Science. 305:525–528.

Seixas S, Ferrand N, Rocha J. 2002. Microsatellite variation and evolution of the human Duffy blood group polymorphism. Mol Biol Evol. 19:1802–1806.

Seixas S, Garcia O, Trovoada MJ, Santos MT, Amorim A, Rocha J. 2001. Patterns of haplotype diversity within the serpin gene cluster at 14q32.1: insights into the natural history of the alpha1-antitrypsin polymorphism. Hum Genet. 108:20–30.

Serre D, Nadon R, Hudson TJ. 2005. Large-scale recombination rate patterns are conserved among human populations. Genome Res. 15:1547–1552.

Shapiro L, Pott GB, Ralston AH. 2001. Alpha-1-antitrypsin inhibits human immunodeficiency virus type 1. FASEB J. 15: 115–122.

Sharp AJ, Locke DP, McGrath SD, et al. (14 co-authors). 2005. Segmental duplications and copy-number variation in the human genome. Am J Hum Genet. 77:78–88.

Smith CL, Power SG, Hammond GL. 1992. A Leu–His substitution at residue 93 in human corticosteroid binding globulin results in reduced affinity for cortisol. J Steroid Biochem Mol Biol. 42:671–676.

Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. Mol Biol Evol. 22:63–73.

Stein PE, Carrell RW. 1995. What do dysfunctional serpins tell us about molecular mobility and disease? Nat Struct Biol. 2: 96–113.

Stephens M, Donnelly P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet. 73:1162–1169.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 68:978–989.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.

Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? Genome Res. 16:702–712.

Tishkoff SA, Varkonyi R, Cahinhinan N, et al. (17 co-authors). 2001. Haplotype diversity and linkage disequilibrium at human

G6PD: recent origin of alleles that confer malarial resistance. Science. 293:455–462.

Tomas G, Seco L, Seixas S, Faustino P, Lavinha J, Rocha J. 2002. The peopling of Sao Tome (Gulf of Guinea): origins of slave settlers and admixture with the Portuguese. Hum Biol. 74: 397–411.

Torpy DJ, Bachmann AW, Gartside M, Grice JE, Harris JM, Clifton P, Easteal S, Jackson RV, Whitworth JA. 2004. Association between chronic fatigue syndrome and the corticosteroid-binding globulin gene ALA SER224 polymorphism. Endocr Res. 30:417–429.

Tuzun E, Sharp AJ, Bailey JA, et al. (12 co-authors). 2005. Fine-scale structural variation of the human genome. Nat Genet. 37:727–732.

van Gent D, Sharp P, Morgan K, Kalsheker N. 2003. Serpins: structure, function and molecular evolution. Int J Biochem Cell Biol. 35:1536–1547.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci. 102:18508–18513.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol. 4:e72.

Wall JD, Przeworski M. 2000. When did the human population size start increasing? Genetics. 155:1865–1874.

Wang X, Grus WE, Zhang J. 2006. Gene losses during human origins. PLoS Biol. 4:e52.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7: 256–276.

[WHO] World Health Organization. 1997. Alpha 1-antitrypsin deficiency: memorandum from a WHO meeting. Bull W H O. 75:397–415.

Xue Y, Daly A, Yngvadottir B, et al. (14 co-authors). 2006. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am J Hum Genet. 78:659–670.

Michael Nachman, Associate Editor

Accepted November 27, 2006

Exhibit "G"

S100A7, S100A10, and S100A11 Are Transglutaminase Substrates[†]

Monica Ruse,[‡] Adam Lambert,[‡] Nancy Robinson,[‡] David Ryan,[‡] Ki-Joon Shon,[‡] and Richard L. Eckert*[,‡,§]

Departments of Physiology and Biophysics, Oncology, Dermatology, Reproductive Biology, and Biochemistry, Case Western Reserve University School of Medicine, 2109 Adelbert Road, Cleveland, Ohio 44106-4970

Received August 18, 2000; Revised Manuscript Received December 14, 2000

ABSTRACT: S100 proteins are a family of 10−14 kDa EF-hand-containing calcium binding proteins that function to transmit calcium-dependent cell regulatory signals. S100 proteins have no intrinsic enzyme activity but bind in a calcium-dependent manner to target proteins to modulate target protein function. Transglutaminases are enzymes that catalyze the formation of covalent ϵ-(γ-glutamyl)lysine bonds between protein-bound glutamine and lysine residues. In the present study we show that transglutaminase-dependent covalent modification is a property shared by several S100 proteins and that both type I and type II transglutaminases can modify S100 proteins. We further show that the reactive regions are at the solvent-exposed amino- and carboxyl-terminal ends of the protein, regions that specify S100 protein function. We suggest that transglutaminase-dependent modification is a general mechanism designed to regulate S100 protein function.

Calcium-dependent signal transduction is a complex process that is mediated by calcium binding proteins. The EF-hand-containing proteins represent an important class of calcium-dependent signal transduction mediator (*1*). These proteins contain one or more EF-hand motifs (*1*). S100 proteins comprise a multigene family that are members of this class of protein. There are at least ten S100 family members, each characterized by a high degree of sequence and structural homology (*2*). S100 proteins are thought to exist in cells as homo- or heterodimers that have no intrinsic enzymatic activity of their own (*1*). In response to calcium binding to the EF-hand structures, S100 proteins undergo a conformational change (*3*−*6*). This conformation change exposes regions of the protein which permit the dimer to bind to and alter the activity of specific target proteins. S100 proteins have been shown to participate in a variety of signal transduction pathways (*4*), often in a cell type-specific manner. S100 proteins regulate cell differentiation, cell cycle progression, energy metabolism, kinase activity, and cytoskeletal membrane interactions, and gain or loss of S100 protein expression has been linked to disease states (*1*).

Transglutaminases are calcium-dependent enzymes that function to form interprotein ϵ-(γ-glutamyl)lysine bonds (*7*). Among other proposed functions, type II transglutaminase functions in apoptosis and type I transglutaminase (keratinocyte transglutaminase) is responsible for cornified envelop formation in surface epithelial cells (*7*). Transglutaminases also function as signal transduction proteins (*8*). For example, TG alters the function of CD38, an enzyme that catalyzes the formation of the second messenger cADPR,[1] by covalently cross-linking it to an as yet unidentified 190 kDa protein (*9*). Transglutaminase-dependent cross-linking also modifies epidermal growth factor receptor function (*10*). In addition, type II transglutaminase has been reported to function as a G protein (*8*).

We previously showed that S100A11 is a transglutaminase substrate (*11*, *12*). However, it has been an open issue whether such reactivity would be observed for other S100 proteins and, if other S100 proteins are reactive, whether they would serve equally efficiently as transglutaminase substrates. In this brief report, we show that transglutaminase can covalently modify several S100 proteins and that individual S100 proteins differ markedly in their TG reactivity. We further show that, in each case, the transglutaminase-reactive amino acid residues are located within the solvent-exposed amino- and carboxy-terminal ends of the proteins and that transglutaminases types I and II share, as a substrate, a common glutamine residue in S100A11. We also provide evidence suggesting that S100A7 may be cross-linked in in vivo.

MATERIALS AND METHODS

Preparation of Transglutaminase. Normal human keratinocytes were grown in twenty-five 75 cm^2 dishes on lethally irradiated 3T3 cells in Dulbecco's modified Eagle's medium/

[†] This work was supported by a grant from the National Institutes of Health (R.L.E.) and utilized the facilities of the Skin Diseases Research Center of Northeast Ohio (AR39750). N.R. was supported as a trainee of the National Institutes of Health Training Program in Cardiovascular Research (HL07653). A.L. was a trainee on the National Institutes of Health Training Program in Aging. Peptide sequencing, performed at the Case Western Reserve University Molecular Biology Core Laboratory, was supported by National Institutes of Health Grant P30 CA443703.

* To whom correspondence should be addressed. Phone: 216-368-5530. Fax: 216-368-5586. E-mail: rle2@po.cwru.edu.

[‡] Department of Physiology and Biophysics.

[§] Departments of Oncology, Dermatology, Reproductive Biology, and Biochemistry.

[1] Abbreviations: cADPR, cyclic ADP-ribose; HEPES, *N*-(2-hydroxyethyl)piperazine-*N'*-2-ethanesulfonic acid; DTT, dithiothreitol; EDTA, ethylenediaminetetraacetic acid; TBS, 10 mM Tris-HCl and 150 mM NaCl, pH 7.4; HPLC, high-pressure liquid chromatography; TFA, trifluoroacetic acid; TG1, type I transglutaminase; TG2, type II transglutaminase; E-FABP, epidermal fatty acid binding protein; HRP, horseradish peroxidase.

10.1021/bi0019747 CCC: $20.00
Published on Web 02/15/2001

F-12 (3:1) containing 5% delipidized fetal calf serum (*13, 14*), 100 μM nonessential amino acids, 100 units/mL penicillin, 100 mg/mL streptomycin, 5 μg/mL transferrin, 2 nM T3, and 0.1 nM cholera toxin. At 6 days post-confluence, the cells were washed with phosphate-buffered saline and scraped into homogenization buffer (10 mM HEPES, pH 7.4, 2 mM EDTA, and 2 mM DTT at 1 mL/dish). The cells were lysed by Dounce homogenization (B-pestle) and centrifuged at 100000*g* for 30 min. The pellet was resuspended in 25 mL of homogenization buffer containing 0.2% Triton X-100, homogenized (Dounce, B-pestle), and maintained on ice for 30 min. The extract was centrifuged at 100000*g* for 30 min, and the supernatant containing transglutaminase type I (TG1 activity) was stored as 0.5 mL aliquots at −80 °C. The activity of the preparation was 0.032 unit/mL when measured by monitoring [^{14}C]putrescine incorporation into dimethylcasein (1 unit = 1 μmol of putrescine incorporated/h). Transglutaminase type II (TG2, porcine liver, 1.6 units/mg, Sigma) was purchased from Sigma Chemicals, reconstituted in sterile distilled H_2O, and stored at −80 °C. TG1 and TG2 activities were normalized on the basis of the ability to [^{14}C]-putrescine-label dimethylcasein (Sigma) (*15*).

Production and Purification of Recombinant S100 Proteins. cDNAs for S100 proteins were produced from polyadenylated RNA isolated from normal human keratinocytes. S100 protein-specific primers containing *Nde*I (upstream) and *Bam*HI (downstream) restriction sites were used to amplify S100A10, S100A11, and S100A7 by polymerase chain reaction [S100A7 (upstream primer) 5′-GGC ATA TGA GCA ACA CTC AAG CTG AG, (downstream primer) 5′-TAG GAT CCT GGG TCT CTG GAG GCC CAT TG; S100A10, (upstream primer) 5′-GGC ATA TGC CAT CTC AAA TGG AAC ACG CC, (downstream primer) 5′-TAG GAT CCT TAT CAG GGA GGA GCG AAC TGC; S100A11, (upstream primer) 5′-CAT ATG GCA AAA ATC TCC AGC CC, (downstream primer) 5′-GGA TCC TGA GGT GGT TAG TGT GCT CA] (*11, 16, 17*). The S100 protein-encoding cDNAs were then cloned into the pET28a+ bacterial expression vector (Novagen). S100 proteins were expressed in *Escherichia coli* bacterial strain BL21-DE3 transformed with the pET28a+-S100 expression vector using IPTG to induce expression (*18*). The recombinant S100 proteins contains a polyhistidine sequence at the amino terminus that was used for isolation to homogeneity using immobilized metal affinity chromatography. The polyhistidine tract was subsequently removed by thrombin cleavage. Following thrombin cleavage to remove the polyhistidine track, each recombinant S100 protein retains a three amino acid extension (Gly-Ser-His) at the amino terminus.

Synthesis of Biotinylated Amine Acceptor Peptide. An amine acceptor hexapeptide, TVQQGL, was synthesized. The hexapeptide was then biotinylated by incubation with 4 mg of NHS-LC-biotin (Pierce) for 2 h at room temperature in 1 mL of 0.1 M NH_4HCO_3, pH 8. The reaction was terminated by adding 50 μL of 1 M Tris-HCl, pH 7.5, and the product was lyophilized and dissolved in water at a final concentration of 100 mM (*19, 20*).

In Vitro Cross-Linking of Recombinant S100 Proteins. Recombinant human S100 protein (25 μM) was reacted in vitro, in the presence or absence of [^{14}C]putrescine ([^{14}C]-putrescine, 20 μCi, 0.7 mM; Amersham) as previously described (*11*), or with 2 mM amine acceptor hexapeptide (*20*), in 100 μL reactions containing 50 mM Tris-HCl, pH 7.4, 1 mM EDTA, 3.4 mM DTT, 30 mM NaCl, 0.1% Triton X-100, and 0.002−0.067 unit of type II (guinea pig liver) or type I (keratinocyte) transglutaminase. The activity of each enzyme was normalized on the basis of the ability to incorporate [^{14}C]putrescine into dimethylcasein (*15*). Reactions were initiated by the addition of 10 mM $CaCl_2$, incubated at 37 °C for 1−3 h, and terminated by addition of EDTA to 20 mM on ice. Samples were analyzed by denaturing polyacrylamide gel electrophoresis and/or trichloroacetic acid precipitation onto Millipore type HA filters for scintillation counting. The remainder was dialyzed overnight against Tris-buffered saline (TBS, 10 mM Tris-HCl, 150 mM NaCl, pH 7.4) to remove unreacted putrescine prior to enzymatic digestion.

HPLC Separation and Microsequencing of Proteolytic Peptides. Recombinant human S100 protein was reacted in vitro in the presence of [^{14}C]putrescine as described above. Cross-linked S100A10 and S100A11 were digested with V8 protease (Worthington) and trypsin (Worthington), respectively, as previously described (*11*). The resulting protein fragments were separated by reverse-phase HPLC using a Waters 600E system equipped with a Waters 484 absorbance detector. Samples were acidified to 0.2% (v/v) with trifluoroacetic acid (TFA) and loaded on a C18 reverse-phase column (Advantage-100, 5 mm, 240 × 4.6 mm; Thompson Liquid Chromatography, Springfield, VA) equilibrated with 0.1% TFA at a flow rate of 1 mL/min. After 15 min, a linear acetonitrile gradient was started and increased at a rate of 0.6%/min. Purified peptides were assayed for radioactivity by scintillation counting. Peaks containing radioactivity were dried, resuspended in 70% acetonitrile (0.1% TFA), spotted to BioBrene Plus-treated (Applied Biosystems) glass fiber filters, and sequenced using a Perkin-Elmer/Applied Biosystems Procise model 494 microsequinator. For each peptide two sequencing runs were completed. The first run determined the sequence of the purified peptide. During the second sequencing run individual cycles were collected and assayed for [^{14}C]putrescine-dependent radioactivity.

Gel Electrophoretic Methods. Samples were boiled in Laemmli sample buffer containing 2% β-mercaptoethanol and electrophoresed on 16% polyacrylamide gels. For gels containing radioactivity, proteins were stained with Coomassie blue, and gels were soaked in Fluor-Hance (RPI, Mount Prospect, IL), dried, and exposed to X-ray film. For immunoblot analysis, proteins were transferred to nitrocellulose membrane. The membrane was blocked with 5% milk, incubated with the appropriate rabbit 1° antisera (anti-S100A7, anti-S100A10, anti-S100A11, or preimmune) at a 1:2500 dilution. The membranes were washed and incubated with a horseradish peroxidase-conjugated goat anti-rabbit IgG secondary antibody (Amersham Corp., 1:10 000 dilution), and antibody binding was visualized using chemiluminesence detection reagents (Amersham Corp.) (*11*). Detection of biotinylated hexapeptide (TVQQGL) incorporation was monitored using horseradish peroxidase conjugated to streptavidin reagent (Vector labs) diluted 1:200.

Preparation of Human Psoriasis Tissue. Uninvolved and involved (active psoriatic plague) epidermis was harvested from psoriatic patients using a dermatome and stored at −80 °C. For analysis, samples of the tissue were homogenized, and equivalent quantities of protein were fractionated on a

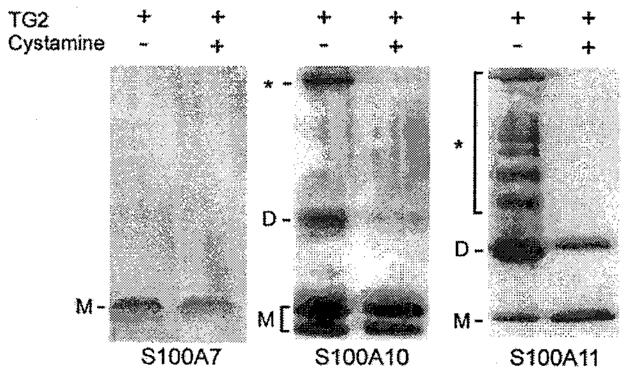


FIGURE 1: S100 proteins as transglutaminase substrates. Equal concentrations (25 μM) of S100A7 (left), S100A10 (center), and S100A11 (right) were incubated with type II transglutaminase (0.005 unit) in the presence or absence of 20 mM cystamine. The products were then electrophoresed on an 16% polyacrylamide gel in denaturing and reducing conditions, transferred to nitrocellulose, and incubated with (left to right) S100A7-, S100A10-, or S100A11-specific antibodies. Binding of the these antibodies was visualized by subsequent incubation with horseradish peroxidase-conjugated goat anti-rabbit IgG, followed by visualization with chemiluminescent reagents. M, D, and *, respectively, indicate the migration of S100 protein monomers, dimers, and multimers. This experiment is representative of three independent experiments. No cross-linking was observed when TG2 was omitted from the reaction mixture.

12% polyacrylamide gel. The presence of S100A7 was monitored by immunoblot using an S100A7-specific antibody that was generated in rabbits in response to recombinant human S100A7 (Robinson and Eckert, unpublished).

RESULTS

S100 Proteins as Transglutaminase Substrates. To study S100 protein transglutaminase reactivity, we incubated recombinant S100 proteins with transglutaminase and calcium and then measured multimer formation by gel electrophoresis and immunoblot. The results from reaction of S100A7, S100A10, and S100A11 with type II transglutaminase (TG2) are shown in Figure 1. Multimer formation is observed for both S100A10 and S100A11 and is inhibited by the transglutaminase inhibitor cystamine. In contrast to the results with S100A11 and S100A10, multimer formation is not detected with S100A7. These results suggest that S100A10 and S100A11 possess both reactive glutamine and lysine residues and that S100A7 is not able to form intermolecular cross-links. No reaction was observed when transglutaminase was omitted from the reaction (not shown).

Evidence for Reactive Glutamine Residues. As a more sensitive method of detecting reactive glutamines, S100A7, S100A10, and S100A11 were incubated with transglutaminase and calcium in the presence of the amine donor [^{14}C]-putrescine (*11*). As shown in Figure 2, all three proteins incorporated [^{14}C]putrescine, indicating that each protein contains TG-reactive glutamine residue(s). S100A10 and S100A11 incorporated [^{14}C]putrescine into monomers, dimers, and multimers, while S100A7 incorporated radioactivity only into monomers. To estimate the relative reactivity of each substrate, equivalent amounts of each S100 protein were incubated with a fixed concentration of TG2 and assayed for [^{14}C]putrescine incorporation. Total incorporation was measured by monitoring radioactivity in trichloroacetic acid-precipitable S100 protein. S100A11 and S100A10 incorporated 6 and 19 times more [^{14}C]putrescine, respectively, compared to S100A7.

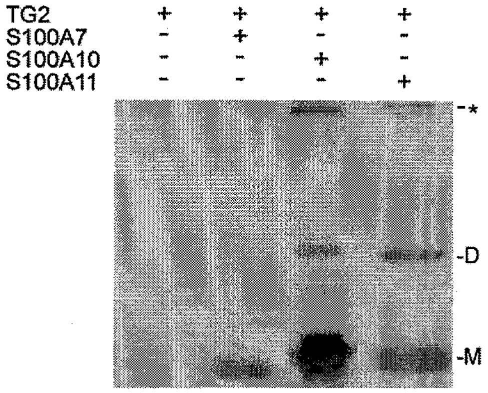


FIGURE 2: [^{14}C]Putrescine incorporation into S100 proteins. Equivalent amounts of S100A7, S100A10, and S100A11 (25 μM) were incubated for 30 min with 880 μM [^{14}C]putrescine, 0.005 unit of TG2, and 20 mM calcium chloride. The resulting cross-linked products were electrophoresed as described in Figure 1, and the gels were fluorographed and exposed on film. The S100A7-, S100A10-, and S100A11-containing lanes were exposed for 4 days, 1 day, and 2 days, respectively. M, D, and *, respectively, indicate migration of monomers, dimers, and multimers. This experiment was repeated three times with similar results. No incorporation of [^{14}C]putrescine was observed when TG2 was omitted from the reaction mixture.

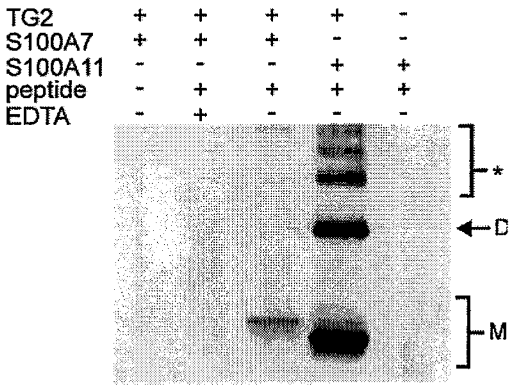


FIGURE 3: Detection of reactive lysines in S100A7. Recombinant human S100A7 (25 μM) or S100A11 (25 μM), 0.005 unit of TG2, and 20 mM $CaCl_2$ were incubated for 60 min at 37 C in the presence or absence of 2 mM biotinylated peptide or 20 mM EDTA. The reaction products were then electrophoresed on a denaturing 16% polyacrylamide gel, transferred to nitrocellulose, and incubated with streptavidin-conjugated horseradish peroxidase (HRP, 5 μg/mL). The blots were washed and incubated to develop the HRP-dependent signal. M, D, and *, respectively, indicate the migration of S100 protein monomers, dimers, and multimers. Similar results were observed in each of four experiments.

Reactive Lysine Residues in S100A7. Since S100A7 encodes a reactive glutamine residue (Figure 2) but does not form multimers (Figure 1), it is possible that S100A7 lacks a reactive lysine residue. To test this possibility, S100A7 protein was incubated with transglutaminase and the biotinylated hexapeptide, TVQQEL, which functions as an amine acceptor (*20*). The reaction products were electrophoresed and transferred to nitrocellulose, and incorporation of the hexapeptide was monitored by streptavidin blot analysis (*19*, *20*). As shown in Figure 3, the amine acceptor peptide was covalently linked to the S100A7 monomer. However, no labeled multimers were detected. The presence of labeled

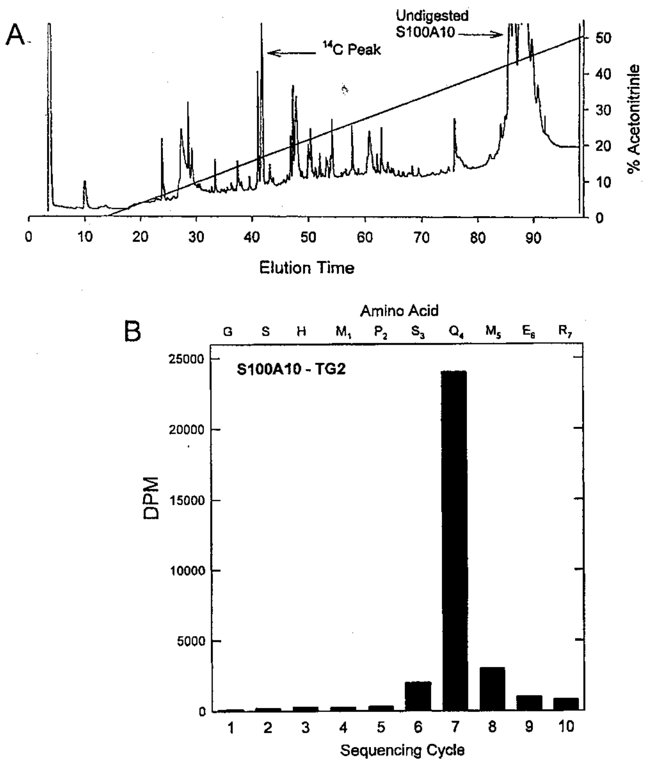


FIGURE 4: Identification of the TG2-reactive glutamine residues in S100A10. S100A10 was cross-linked for 60 min in the presence 20 mM calcium chloride, 880 μM [^{14}C]putrescine, and 0.005 unit of type II transglutaminase. The resulting cross-linked products were V8 protease-digested, and the protease-cleavage products were chromatographed using a reverse-phase HPLC column (A). V8 protease-digested peptides were eluted using an acetonitrile gradient (diagonal line), and peptide elution was detected using a Waters 484 absorbance detector. The radioactivity, which eluted as a single peak (^{14}C peak), was collected for microsequencing. As shown in panel B, the radioactivity was released in cycle 7 corresponding to amino acid Q_4.

monomers suggests that S100A7 contains a reactive lysine residue. In contrast to the low amount of hexapeptide linked to S100A7 monomer, high-level incorporation of hexapeptide was observed into S100A11 monomers, dimers, and multimers (Figure 3).

Identification of TG2-Reactive Glutamine Residues in S100A10 and S100A7. S100A11 is known to contain only two transglutaminase-reactive residues: K_3 at the amino terminus and Q_{102} at the carboxyl terminus (*11*). As S100 proteins share a common structure, it could be expected that TG reactive residues in other S100 proteins would be located within corresponding domains. To evaluate this possibility, we incubated S100A10 with TG2 in the presence of [^{14}C]-putrescine. Following cross-linking the products were digested with V8 protease, and the resulting peptide fragments were separated by HPLC. The single radioactive peak, eluting at 42.5 min, was collected for sequencing (Figure 4A). Microsequencing of this peptide identified labeling at the S100A10 amino terminus. The [^{14}C]putrescine label eluted in cycle 7, corresponding to cleavage of residue Q_4 of the S100A10 protein (Figure 4B).

To identify the reactive glutamine in S100A7, the protein was incubated with [^{14}C]putrescine, TG2, and calcium for 1 h at 37 °C. A single radioactive product was observed upon gel electrophoresis that corresponded to the S100A7 monomer. This product was collected and sequenced. The radioactivity was released in sequencing cycle 8, corresponding to the release of S100A7 amino acid Q_5.

Identification of S100 Proteins as Type I Transglutaminase Substrates. The above studies demonstrate that S100 proteins are substrates for type II (TG2) transglutaminase. Transglutaminase type I (TG1) is a second type of transglutaminase. TG1 is involved in assembly of cross-linked structures during terminal keratinocyte differentiation (*7, 21, 22*). To determine whether S100 proteins are TG1 substrates, we incubated S100A11 with TG1 and measured multimer formation and [^{14}C]putrescine incorporation. As shown in Figure 5, TG1 promotes S100A11 dimer and multimer formation (left panel). Moreover, the reaction is inhibited by the transglutaminase inhibitor, cystamine. The opposite panel shows the TG1-dependent incorporation of the amine donor, [^{14}C]putrescine, and that the incorporation is inhibited by EDTA-dependent chelation of the required transglutaminase cofactor, calcium. To identify the reactive glutamine residue, the TG1-cross-linked S100A11 product was digested with trypsin, and the tryptic fragments were separated by HPLC (not shown). A single radioactive peak was identified. Microsequencing the labeled peptide identified the radioactivity as being associated with residue Q_{102}.

We also examined the TG1 reactivity of S100A10. As shown in Figure 6, incubation of S100A10 with TG1 resulted in a low quantity of dimer and multimer formation, which was inhibited by addition of cystamine. However, the low level of incorporation in this case did not permit sequence determination. S100A7 did not appear to function as a substrate for TG1 in this in vitro system.

S100A7: In Vivo Reactivity. The above studies indicate that S100A7 does not form homomultimers in vitro and has low TG reactivity. However, it is possible that S100A7 may become cross-linked to non-S100 proteins in vivo. S100A7 is markedly overexpressed in psoriatic epidermis (*23, 24*). We therefore harvested involved and uninvolved epidermis from several psoriasis patients and assayed for S100A7 by immunoblot. As shown in Figure 7, S100A7 levels are markedly elevated in actively scaling psoriatic epidermis (involved) as compared to nonscaling epidermis (uninvolved). In addition to the presence of S100A7 monomer, high molecular weight anti-S100A7 immunoreactivity bands are observed in both involved and uninvolved samples.

DISCUSSION

Covalent Modification of S100 Proteins. S100 proteins exist in cells as antiparallel, noncovalently associated homo- and heterodimers (*25, 26*). Dimer formation relies on the interaction of hydrophobic globular domains derived from helices I and IV from each monomer (*27, 28*). This pairing forms a highly stable structure that coordinates zinc and calcium ions (*27, 28*). The assembly of these domains leads to the formation of the target protein binding cleft. In response to increased intracellular calcium, there is a conformational change in the dimeric unit to expose hydrophobic surfaces. Exposure of these surfaces permits interaction of the binding cleft with target protein(s) (*29*). Ultimately, this interaction results in a change in distribution and/or activity of the target protein. Because of the potential importance of covalent modification on the structure of the binding cleft and, as a result, on the ability of S100 proteins

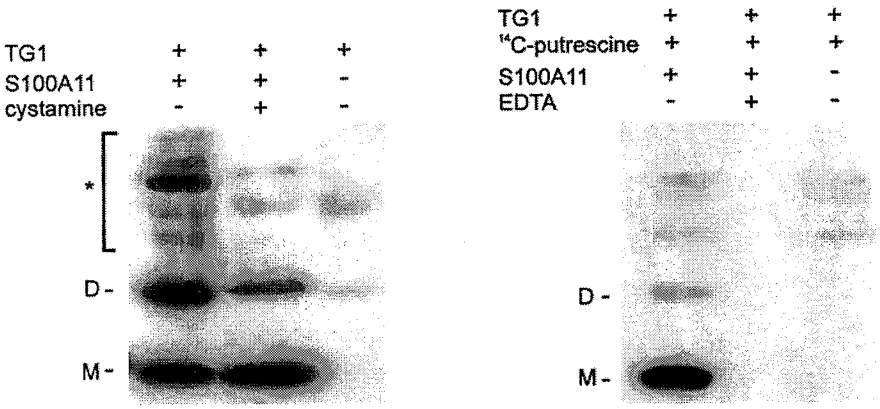


FIGURE 5: S100A11 is a substrate for type I transglutaminase. Recombinant human S100A11 (25 μM) was incubated with type I transglutaminase and calcium in the presence or absence of 20 mM cystamine, 20 mM EDTA, or 880 μM [^{14}C]putrescine for 60 min at 37 °C. The resulting cross-linked products were electrophoresed on a denaturing 16% polyacrylamide gel. Presence of the S100A11 protein was detected by immunoblot using a rabbit anti-human S100A11 antibody (left) or by detecting [^{14}C]putrescine incorporation by fluorography (right). M, D, and *, respectively, indicate the migration of S100 protein monomers, dimers, and multimers. This experiment is representative of four independent experiments.

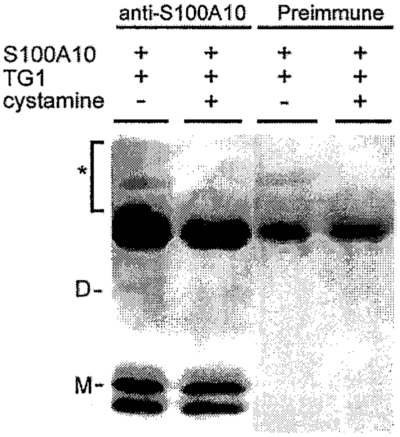


FIGURE 6: TG1-dependent cross-linking of S100A10. Recombinant human S100A10 (25 μM) was incubated with 20 mM calcium chloride and 0.005 unit of TG1 for 60 min at 37 °C in the presence or absence of 20 mM cystamine. The reaction products were then electrophoresed in parallel sets of lanes. One set was incubated with rabbit anti-human S100A10 and the other set with preimmune serum. Binding of the primary antibodies was detected by subsequent incubation with horseradish peroxidase-conjugated goat anti-rabbit IgG secondary antibody. M, D, and *, respectively, indicate migration of S100A10 monomers, dimers, and multimers. Similar results were observed in each of four separate experiments.

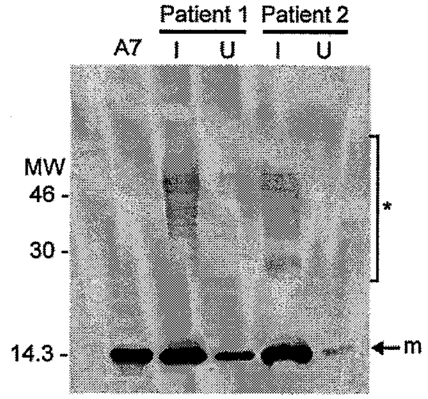


FIGURE 7: Detection of S100A7 in psoriatic epidermis. Epidermis was harvested from psoriasis patients (1 and 2) from both involved (I) and uninvolved (U) regions. Protein extracts were prepared, and the equivalent amount of protein was electrophoresed, transferred to nitrocellulose, and blotted with a rabbit anti-human S100A7 antibody. The letter m indicates migration of S100A7 monomers, while the asterisk indicates migration of higher molecular weight immunoreactive S100A7 bands that may be cross-linked. Recombinant human S100A7 was electrophoresed in the lane labeled A7. This experiment was repeated four times with similar results.

Table 1: Summary—Transglutaminase Modification of S100 Proteins

S100 protein	TG substrate	reactive residues	homo-multimer formation	in vivo reactivity
S100A7	yes (TG1[a] and -2)	Q_5, $K_?$[b]	no	yes (epidermis)[a]
S100A10	yes (TG1 and -2)	Q_4, $K_?$	yes	yes (epidermis) (*11, 12*)
S100A11	yes (TG1 and -2)	Q_{102}, K_3	yes	yes (epidermis) (*11, 12*)

[a] S100A7 appears in psoriatic epidermal extracts as monomers and higher molecular weight bands. We hypothesize that these are cross-linked products; however, this must be directly confirmed. [b] $K_?$ indicates that a reactive lysine is present but that the specific residue is not known.

to regulate target protein function, it is important to identify posttranslational modifications that may modify S100 protein structure. S100A8 and S100A9, for example, are phosphorylated, and this modification is thought to enhance S100-associated calcium binding and association with the plasma membrane (*30*, *31*).

Transglutaminase enzymes catalyze the formation of covalent interprotein ε-(γ-glutamyl)lysine bonds (*32*−*35*). In a previous report, we identified S100A11 as a transglutaminase substrate (*11*, *12*). However, it was not clear that transglutaminase reactivity would be a property shared by other S100 proteins. In the present study we compare the transglutaminase reactivity of S100A10, S100A11, and S100A7. As summarized in Table 1, although all S100 proteins are substrates, each displays substantial differences in reactivity. In the case of S100A10 and S100A11, transglutaminase promotes the formation of covalently linked homomultimers. The major product formed in this reaction is a dimer with lesser quantities of higher multimers. Since each S100 protein appears to contain a single reactive glutamine and a single reactive lysine residue (*11*), we suspect that these homomultimers consist of head-to-tail linked monomers. In contrast, S100A7 does not form homomultimers in vitro, suggesting that, with respect to transglutaminase reactivity, S100A7 is fundamentally different from S100A10 and S100A11. Additional studies, using small molecular probes, indicate that S100A7 does in fact contain transglutaminase-reactive glutamine and lysine residues. The low reactivity of S100A7 may be explained by the sequence of the S100A7 protein. The single lysine in the carboxy-terminal domain, K_{88}, is located relatively close to the central globular EF-hand domain, which may reduce its reactivity. It is interesting that psoriatic epidermis contains putative S100A7 multimers, consistent with the hypothesis that S100A7 functions as a type I transglutaminase substrate in tissue. If these forms are cross-linked, it would suggest that S100A7 does not cross-link to itself but does become cross-linked to non-S100 proteins. A candidate for a cross-linking partner is epidermal fatty acid binding protein (E-FABP). E-FABP has been reported to form a noncovalent complex with S100A7 (*36*, *37*). Additional studies will be necessary to determine whether S100A7 and E-FABP become covalently associated via a transglutaminase-dependent mechanism.

Location of Transglutaminase-Reactive Amino Acids. As noted above, S100 proteins share a common structure (*29*). Thus, it was of interest to identify the TG-reactive amino acids. These include Q_5 in S100A7, Q_4 in S100A10, and Q_{102} and K_3 in S100A11. It is noteworthy that all of these residues are localized in the solvent-exposed amino- and carboxy-terminal S100 protein flanking domains. This is consistent with the idea that transglutaminases prefer to modify highly accessible solvent-exposed glutamine and lysine residues (*7*, *38*−*41*). It has important functional implications that these residues are located within regions that are required for S100 protein interaction with target proteins (*29*). Since the carboxyl terminus of each S100 protein is located fairly close to the homodimeric fold (target protein binding site) and since a conformational change is required for target protein binding, we postulate that transglutaminase-dependent covalent modification at the carboxyl terminus will cause inappropriate S100 protein function. Thus, we postulate that transglutaminase-dependent modification may be a general mechanism designed to inactivate S100 protein function.

Transglutaminase exists in several forms (*7*, *21*, *22*, *42*). Each form is expressed in specific tissues and appears to have a distinct function. TG1, for example, is expressed in keratinocytes and plays a major role in assembling the keratinocyte cornified envelop (*43*, *44*). TG3 also has a role in covalent assembly of structures during keratinocyte differentiation (*42*). TG2 (tissue-type transglutaminase) is a ubiquitously expressed, soluble enzyme that functions to regulate receptor function (*8*, *10*), apoptosis (*45*, *46*), cross-linking of extracellular matrix (*47*), and other cellular processes (*8*, *10*). The above results show that S100 proteins are both TG1 and TG2 substrates. In addition, our results suggest that both TG1 and TG2 modify the same sites on S100A11 (i.e., Q_{102}) and the rank order of reactivity of the three S100 proteins is similar regardless of which TG is involved. On the basis of these results, we hypothesize that TG-dependent modification of S100 proteins may be a general mechanism that terminates S100 action. Studies are currently underway to test this hypothesis.

REFERENCES

1. Donato, R. (1999) *Biochim. Biophys. Acta 1450*, 191−231.
2. Donato, R. (1990) *Adv. Exp. Med. Biol. 269*, 103−106.
3. Smith, S. P., and Shaw, G. S. (1998) *Structure 6*, 211−222.
4. Rustandi, R. R., Baldisseri, D. M., Drohat, A. C., and Weber, D. J. (1999) *Protein Sci. 8*, 1743−1751.
5. Drohat, A. C., Baldisseri, D. M., Rustandi, R. R., Weber, D. J., and Wilder, P. T. (1998) *Biochemistry 37*, 2729−2740.
6. Sastry, M., Ketchem, R. R., Crescenzi, O., Weber, C., Lubienski, M. J., Hidaka, H., and Chazin, W. J. (1998) *Structure 6*, 223−231.
7. Greenberg, C. S., Birckbichler, P. J., and Rice, R. H. (1991) *FASEB J. 5*, 3071−3077.
8. Zhang, J., Tucholski, J., Lesort, M., Jope, R. S., and Johnson, G. V. (1999) *Biochem. J. 343 (*Part 3), 541−549.
9. Umar, S., Malavasi, F., and Mehta, K. (1996) *J. Biol. Chem. 271*, 15922−15927.
10. Katoh, S., Hashimoto, M., Kohno, H., and Ohkubo, Y. (1993) *Arch. Biochem. Biophys. 303*, 421−428.
11. Robinson, N. A., and Eckert, R. L. (1998) *J. Biol. Chem. 273*, 2721−2728.
12. Robinson, N. A., Lapic, S., Welter, J. F., and Eckert, R. L. (1997) *J. Biol. Chem. 272*, 12035−12046.
13. Gilfix, B. M., and Eckert, R. L. (1985) *J. Biol. Chem. 260*, 14026−14029.
14. Rheinwald, J. G., and Green, H. (1975) *Cell 6*, 331−343.
15. Lambert, A., Ekambaram, M. C., Robinson, N. A., and Eckert, R. L. (2000) *Skin Pharmacol. Appl. Skin Physiol. 13*, 17−30.
16. Madsen, P., Rasmussen, H. H., Leffers, H., Honore, B., Dejgaard, K., Olsen, E., Kiil, J., Walbum, E., Andersen, A. H., Basse, B., et al. (1991) *J. Invest. Dermatol. 97*, 701−712.
17. Dooley, T. P., Weiland, K. L., and Simon, M. (1992) *Genomics 13*, 866−868.
18. LaCelle, P. T., Lambert, A., Ekambaram, M. C., Robinson, N. A., and Eckert, R. L. (1998) *Skin Pharmacol. Appl. Skin Physiol. 11*, 214−226.
19. Groenen, P. J., Smulders, R. H., Peters, R. F., Grootjans, J. J., van den Ijssel, P. R., Bloemendal, H., and de Jong, W. W. (1994) *Eur. J. Biochem. 220*, 795−799.
20. Groenen, P. J., Bloemendal, H., and de Jong, W. W. (1992) *Eur. J. Biochem. 205*, 671−674.
21. Kim, I. G., McBride, O. W., Wang, M., Kim, S. Y., Idler, W. W., and Steinert, P. M. (1992) *J. Biol. Chem. 267*, 7710−7717.
22. Phillips, M. A., Qin, Q., Mehrpouyan, M., and Rice, R. H. (1993) *Biochemistry 32*, 11057−11063.
23. Algermissen, B., Sitzmann, J., LeMotte, P., and Czarnetzki, B. (1996) *Arch. Dermatol. Res. 288*, 426−430.
24. Tavakkol, A., Zouboulis, C. C., Duell, E. A., and Voorhees, J. J. (1994) *Mol. Biol. Rep. 20*, 75−83.
25. Hilt, D. C., and Kligman, D. (1991) The S-100 protein family: a biochemical and functional overview, in *Novel calcium-binding proteins: Fundamentals and clinical implications* (Heizmann, C. W., Ed.) Springer-Verlag, Berlin.
26. Schafer, B. W., and Heizmann, C. W. (1996) *Trends. Biochem. Sci. 21*, 134−140.
27. Rety, S., Osterloh, D., Arie, J. P., Tabaries, S., Seeman, J., Russo-Marie, F., Gerke, V., and Lewit-Bentley, A. (2000) *Struct. Folding Des. 8*, 175−184.
28. Rety, S., Sopkova, J., Renouard, M., Osterloh, D., Gerke, V., Tabaries, S., Russo-Marie, F., and Lewit-Bentley, A. (1999) *Nat. Struct. Biol. 6*, 89−95.
29. Donato, R. (1986) *Cell Calcium 7*, 123−145.
30. van den Bos, C., Roth, J., Koch, H. G., Hartmann, M., and Sorg, C. (1996) *J. Immunol. 156*, 1247−1254.
31. Guignard, F., Mauel, J., and Markert, M. (1996) *Eur. J. Biochem. 241*, 265−271.
32. Aeschlimann, D., and Thomazy, V. (2000) *Connect. Tissue Res. 41*, 1−27.
33. Autuori, F., Farrace, M. G., Oliverio, S., Piredda, L., and Piacentini, M. (1998) *Adv. Biochem. Eng. Biotechnol. 62*, 129−136.
34. Melino, G., and Piacentini, M. (1998) *FEBS Lett. 430*, 59−63.

35. Nemes, Z., and Steinert, P. M. (1999) *Exp. Mol. Med. 31*, 5−19.
36. Hagens, G., Roulin, K., Hotz, R., Saurat, J. H., Hellman, U., and Siegenthaler, G. (1999) *Mol. Cell. Biochem. 192*, 123−128.
37. Hagens, G., Masouye, I., Augsburger, E., Hotz, R., Saurat, J. H., and Siegenthaler, G. (1999) *Biochem. J. 339* (Part 2), 419−427.
38. Gorman, J. J., and Folk, J. E. (1980) *J. Biol. Chem. 255*, 419−427.
39. Gorman, J. J., and Folk, J. E. (1981) *J. Biol. Chem. 256*, 2712−2715.
40. Gorman, J. J., and Folk, J. E. (1984) *J. Biol. Chem. 259*, 9007−9010.
41. Simon, M., and Green, H. (1988) *J. Biol. Chem. 263*, 18093−18098.
42. Kim, I. G., Gorman, J. J., Park, S. C., Chung, S. I., and Steinert, P. M. (1993) *J. Biol. Chem. 268*, 12682−12690.
43. Eckert, R. L., Crish, J. F., and Robinson, N. A. (1997) *Physiol. Rev. 77*, 397−424.
44. Steven, A. C., and Steinert, P. M. (1994) *J. Cell Sci. 107*, 693−700.
45. Rittmaster, R. S., Thomas, L. N., Wright, A. S., Murray, S. K., Carlson, K., Douglas, R. C., Yung, J., Messieh, M., Bell, D., and Lazier, C. B. (1999) *J. Urol. 162*, 2165−2169.
46. Oliverio, S., Amendola, A., Rodolfo, C., Spinedi, A., and Piacentini, M. (1999) *J. Biol. Chem. 274*, 34123−34128.
47. Haroon, Z. A., Hettasch, J. M., Lai, T. S., Dewhirst, M. W., and Greenberg, C. S. (1999) *FASEB J. 13*, 1787−1795.

BI0019747

Exhibit "H"

Biochemical and Biophysical Research Communications **263,** 135–138 (1999)
Article ID bbrc.1999.1319, available online at http://www.idealibrary.com on IDEAL®

Human S100A11 Exhibits Differential Steady-State RNA Levels in Various Tissues and a Distinct Subcellular Localization

Hiroyasu Inada,* Michiko Naka,* Toshio Tanaka,* Gabriela E. Davey,† and Claus W. Heizmann†,[1]
**Department of Molecular and Cellular Pharmacology, Mie University School of Medicine, Tsu, Mie 514-8507, Japan; and*
†Department of Pediatrics, Division of Clinical Chemistry and Biochemistry, University of Zurich,
Steinwiesstrasse 75, CH-8032 Zurich, Switzerland

Received July 25, 1999

In order to analyze the steady-state RNA levels of S100A11 in different tissues, a cDNA fragment of human S100A11 was isolated from a cDNA library. The obtained fragment was labeled and hybridized to RNA isolated from various tissues. The Northern blot analysis revealed that S100A11 RNA levels varied from high in placenta, through intermediate in heart, lung, kidney, and most muscle samples, to barely detectable in brain. An efficient purification method for recombinant S100A11 yielding high quantities was developed. Furthermore, to examine the subcellular localization of this protein, the human polypeptide S100A11 antibodies were raised in rabbit. S100A11 was found to have a localization distinct from other S100 proteins examined, and is mostly localized in the nucleus, with slight variations among different glioblastoma cell types. © 1999 Academic Press

Ca^{2+}-binding S100 proteins became of major interest because of their close association with several human diseases (1, 2) and their use as prognostic markers in different tumor types (3, 4). S100A11 (previously named S100C or calgizzarin) is a novel and less known member of this large protein family, exhibiting several unique properties. S100A11, first purified and partially characterized from porcine heart (5, 6) and chicken gizzard (7), was subsequently found to be highly expressed in colorectal cancer (8). The low expression in normal colon tissue suggested a role in cell transformation. This would be consistent with the localization of the S100A11 gene on human chromosome 1q21 (9, 10), a region most frequently amplified in tumor tissues. S100A11 was found to regulate cytoskeletal function via Ca^{2+}-dependent interaction with annexin I (5, 11), targeting S100A11 to early endosomes (12). Recently, S100A11 was discovered to be a component of the keratinocyte cornified envelope (13) post-translationally modified by transglutaminase. The purpose of this study was to investigate the S100A11 RNA steady-state levels in a number of human tissues and to determine the distinct intracellular localization of S100A11 by confocal laser scanning microscopy.

MATERIALS AND METHODS

Screening of a human cDNA library. A cDNA fragment of human S100A11 was amplified from a human cDNA library (Stratagene) by the polymerase chain reaction (PCR). The oligonucleotide primers for PCR were based on the 5′- and 3′-ends of the open reading frame of the porcine S100A11 cDNA (5). The nucleotide sequence was determined using the automated DNA Sequencer (ABI, model 377). Database searches in GenBank and sequence analysis were performed by the DNA Data Bank of Japan (DDBJ, National Institute of Genetics, Mishima, Japan).

Northern blot analysis. The Northern blot membrane containing human poly(A)+ RNA from various tissues was purchased from Clontech. The filter was hybridized with a human S100A11 cDNA open-reading frame fragment labeled with [γ-^{32}P]dCTP (Amersham), according to a Megaprime protocol (Amersham). Hybridization was carried out at 68°C in ExpressHyb buffer (Clontech, according to the manufacturer's instructions). The membrane was exposed to X-ray film (Kodak) for 2 or 16 h at −80°C.

Expression, purification of GST human S100A11 fusion protein, production of anti-human S100A11 peptide polyclonal antibody, and Western blot analysis. A cDNA fragment containing the open reading frame of the human S100A11 coding sequence was cloned into the BamHI and the EcoRI sites of the pGEX-2T (Pharmacia) expression vector. The GST human S100A11 protein was expressed in *E. coli* NM522 and lysates were prepared as described earlier (14). One liter of culture was incubated at 37°C until the absorbance at 600 nm reached 0.7. At this point the culture was supplemented with IPTG (1 mM final concentration) and incubated for additional 20 h at 25°C. The cells were centrifuged at 6000 *g* for 10 min. The pellets were resuspended in 30 ml of TBS (25 mM Tris-HCl, 135 mM NaCl, pH 7.4) containing 1% Triton X-100 (v/v). The recombinant protein was

Nomenclature of S100 proteins: Schäfer *et al.* (1995) *Genomics* **25,** 638–643; Wicki *et al.* (1996) *Cell Calcium* **20,** 459–464; Wicki *et al.* (1996) *Biochem. Biophys. Res. Commun.* **227,** 594–599.

[1] Corresponding author. Fax: +4112667169. E-mail: heizmann@kispi.unizh.ch.

0006-291X/99 $30.00

purified by glutathione Sepharose 4B column chromatography (Pharmacia).

A peptide corresponding to 92-105 amino acids of human S100A11 (HDSFLKAVPSQKRT) was synthesized. Antibodies against human S100A11 peptide were raised in rabbits through series of injections containing the peptides and Freund's complete adjuvant (15). In order to analyze the efficiency of the purification protocol and the purity of the S100A11 recombinant protein, different fractions from the purification process plus the final product were electrophoresed on 12.5% SDS-polyacrylamide gel. Proteins were stained with Coomassie Blue or electroeluted onto nitrocellulose membranes, immunoblotted with anti-S100A11 followed by incubation with horseradish peroxidase-conjugated anti-rabbit IgG, and visualized by enhanced chemiluminescence.

Cell culture and immunofluorescent staining. Human glioblastoma cell lines: U-373MG, U-87MG (HTB-17, HTB-14, respectively, available from ATCC), LN215, and LN444 (a generous gift from Erwin G. Van Meir) were maintained in 10% fetal calf serum (FCS) in Dulbecco's modified Eagle's medium (DMEM) with penicillin-streptomycin. In order to stain the cells, they were rinsed with DMEM and fixed with 3.7% formaldehyde in DMEM. Cells were permeabilized with methanol, washed in 5% horse serum (HS)-DMEM, and incubated with specific antibodies (anti-human S100A11, described above, dilution 1:500, and anti-human S100A6 raised in goat, described in (18), dilution 1:1000) for 1 h at 37°C. After washing the cells twice in 5% HS-DMEM, they were incubated with secondary Cy3-conjugated antibodies (goat anti-rabbit IgG (H + L) or mouse anti-goat IgG (H+L), respectively, Jackson ImmunoResearch Laboratories, Inc.) for 45 min at 37°C. The cells were washed twice in 5% HS-DMEM, once in PBS (pH 9), and mounted in Molwiol (Hoechst) containing 0.75% n-propyl-gallate as an anti-bleaching agent. Mounted slides were left to dry for 24 h at room temperature in the dark, and stored in the dark at 4°C until viewed. As a control, the cells were incubated with pre-immune sera.

Confocal laser scanning microscopy. Cells were scanned with a Zeiss Axioplan fluorescence microscope (100× oil objective) equipped with a confocal unit MRC-600 (Biorad) and an argon-krypton laser with an excitation wavelength of 568 nm, and a long pass filter of 585 nm. Subsequently, the images were processed using Imaris (Bitplane) and Photoshop (Adobe System) software on a SGI-Indigo 2 workstation (Silicon Graphics) described in (16, 17).

RESULTS AND DISCUSSION

The glutathione Sepharose 4B column chromatography yields high quantity of pure S100A11. In order to purify large quantities of S100A11, the cDNA was subcloned into the pGEX-2T expression vector and expressed in *E. coli.* The GST human S100A11 fusion protein was successfully purified using glutathione Sepharose 4B column chromatography (Fig. 1). The analysis by SDS-PAGE and enhanced chemiluminescence revealed two bands: human S100A11 and GST human S100A11 fusion protein. The antibody against human S100A11 did not cross-react with porcine S100A11 protein.

S100A11 steady-state RNA levels vary between different tissues. Using Northern blot analysis, we examined steady-state RNA levels of human S100A11 in various adult (Figs. 2A and 2B) and some fetal (Fig. 2C) human tissues. A single-species transcript of expected size (about 0.5 kb) was detected at high levels in placenta. Intermediate levels were found in adult heart, adult and fetal lung and kidney, as well as adult pancreas. Low levels were detected in adult skeletal muscle, as well as adult and fetal liver. S100A11 was barely detectable in fetal or adult brain. Further analysis of various adult muscle tissues showed that although S100A11 is present in all of the samples examined, its levels vary greatly from rather low in skeletal muscle and colon to quite high in bladder.

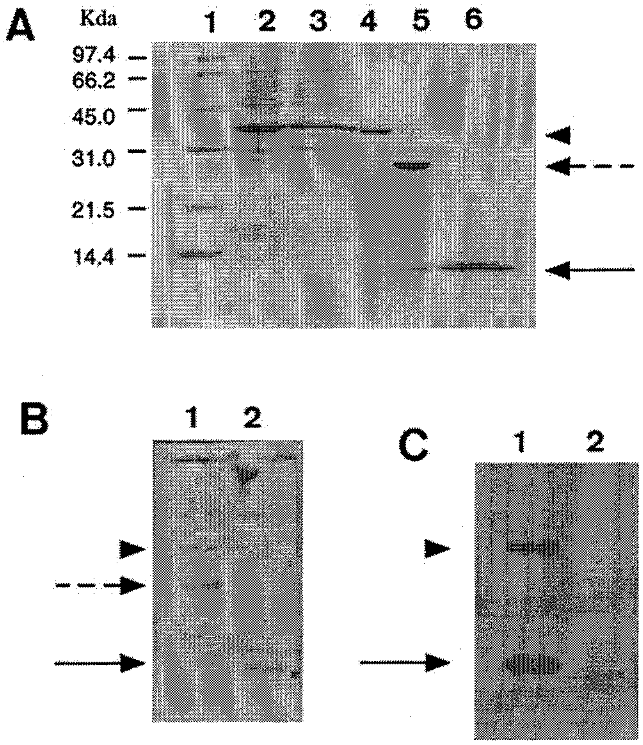


FIG. 1. Expression and purification of recombinant human S100A11. Reducing SDS–polyacrylamide gel (15%) was stained with Coomassie blue. (A) Lane 1, marker; lane 2, crude extract of *E. coli;* lane 3, supernatant fraction of *E. coli* extract; lane 4, eluate from glutathione Sepharose 4B column; lane 5, thrombin cleavage fraction; lane 6, eluate from SP Sepharose fast flow. (B) Lane 1, purified human GST-S100A11 fusion protein; lane 2, porcine S100A11 analyzed by SDS–PAGE (12.5%) and stained with Coomassie blue. (C) Proteins transferred onto nitrocellulose membrane and visualized by enhanced chemiluminescence, using anti-human S100A11 and horseradish peroxidase-conjugated anti-rabbit IgG; lane 1, GST-S100A11 fusion protein; lane 2, porcine S100A11. Arrowhead, GST human S100A11 fusion protein; dotted arrow, GST; arrow, recombinant human S100A11; asterisk, porcine S100A11.

S100A11 localizes mostly in the nucleus of glioblastoma cell lines. To examine the localization of S100A11, the U-373 MG and U-87 MG glioblastoma cells (LN215, LN444, data not shown) were stained with S100A11-specific polyclonal peptide antibody and fluorescent secondary antibody. Figure 3A shows that S100A11 is predominantly localized in the nucleus and less in the cytoplasm of U-373 MG cells. Figure 3B shows localization of S100A11 in the nucleus and in cytoplasm. In contrast, S100A6 (Fig. 3C) localizes

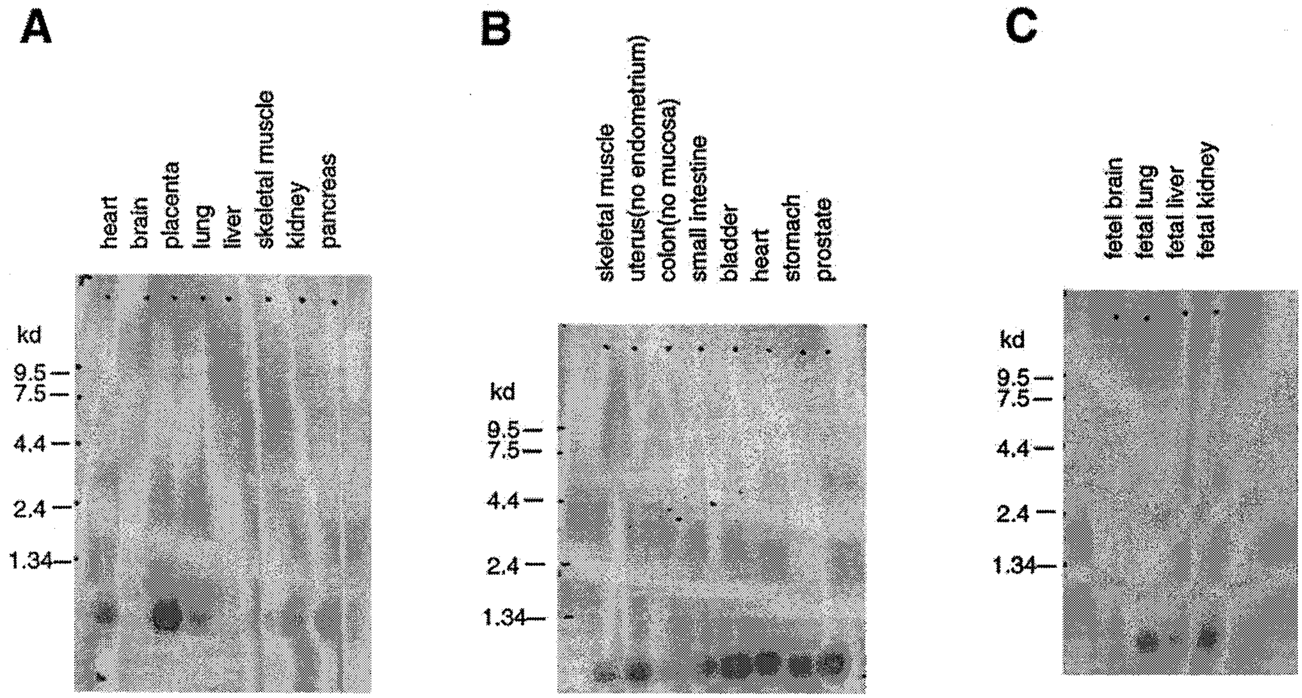


FIG. 2. Steady-state levels of S100A11 mRNA in various human tissues. Commercially available Northern blots were hybridized with the ^{32}P-labeled human S100A11 cDNA fragment. (A) Human multiple-tissue Northern blot. (B) Human muscle multiple-tissue Northern blot. (C) Human fetal multiple-tissue Northern blot.

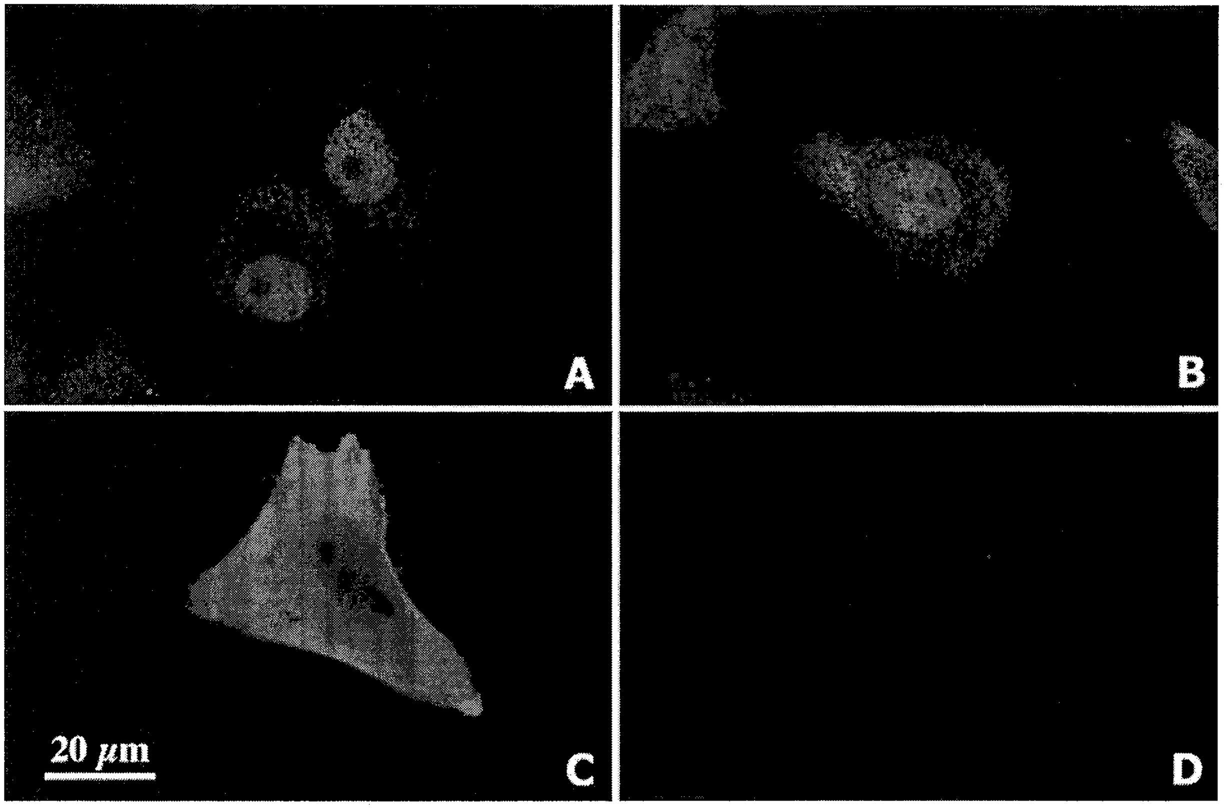


FIG. 3. Subcellular localization. The glioblastoma cells were stained with primary rabbit anti-human S100A11 (dilution 1:500; A, U-373 MG cells; B, U-87 MG cells). (C) Staining with goat anti-human S100A6 (dilution 1:1000). (D) Control staining for S100A11 using corresponding pre-immune serum. All samples were stained with a corresponding secondary Cy3-coupled antibody (anti-rabbit or anti-goat, dilution 1:500).

mostly in the cytoplasm of these cells. These results demonstrate a distinct subcellular localization of S100A11 different from that of other S100 proteins. Furthermore, S100A11, in contrast to other S100 proteins, was found to be expressed at varying levels and locations in different glioblastoma cell lines exhibiting distinct invasiveness and motility rates (data not shown). Considering that S100A11 RNA levels were barely detectable in normal brain tissue, these results suggest that S100A11 might regulate specific cellular processes involved in tumor progression. This is consistent with the preliminary findings suggesting that overexpression of S100A11 in colon cancers might be of importance in the tumor progression process (8).

ACKNOWLEDGMENTS

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, Scientific Research (A)and(B), International Grants Scientific Research (Joint Research) and Exploratory Research from the Ministry of Education, Science, Sports and Culture, and the Grant for Pediatric Diseases from the Ministry of Health and Welfare, Japan. It was also supported in part by grants for Research Project on Muscle Regulation and on Cerebral Vasospasm from Mie University School of Medicine, a Grant-in-Aid (1996) from the Mie Medical Research Foundation, and the Swiss National Science Foundation (31-50510.97). We thank M. Killen for critical reading of the manuscript, as well as Matthias Höchli and Thomas Bächi for their help with confocal microscopy.

REFERENCES

1. Schäfer, B. W., and Heizmann, C. W. (1996) *Trends Biochem. Sci.* **21,** 134–140.
2. Heizmann, C. W., and Cox, J. A. (1998) *BioMetals* **11,** 383–397.
3. Van Ginkel, P. R., Gee, R. L., Walker, T. M., Hu, D.-N., Heizmann, C. W., and Polans, A. S. (1998) *Biochim. Biophys. Acta* **1448,** 290–297.
4. Camby, I., Nagy, N., Lopes, M.-B., Schäfer, B. W., Maurage, C.-A., Ruchoux, M.-M., Murmann, P., Pochet, R., Heizmann, C. W., Brotchi, J., Salmon, I., Kiss, R., and Decaestecker, Ch. (1999) *Brain Pathol.* **9,** 825–843.
5. Naka, M., Qing, Z. X., Sasaki, T., Kise, H., Tawara, I., Hamaguchi, S., and Tanaka, T. (1994) *Biochim. Biophys. Acta* **1223,** 348–353.
6. Ohta, H., Sasaki, T., Naka, M., Hiraoka, O., Miyamoto, C., Furuichi, Y., and Tanaka, T. (1991) *FEBS Lett.* **295,** 93–96.
7. Allen, B. G., Durussel, I., Walsh, M. P., and Cox, J. A. (1996) *Biochem. Cell Biol.* **74,** 687–694.
8. Tanaka, M., Adzuma, K., Iwami, M., Yoshimoto, K., Monden, Y., and Itakura, M. (1995) *Cancer Lett.* **89,** 195–200.
9. Moog-Lutz, Ch., Bouillet, P., Regnier, C. H., Tomasetto, C., Mattei, M.-G., Chenard, M.-P., Anglard, P., Rio, M.-Ch., and Basset, P. (1995) *Int. J. Cancer* **63,** 297–303.
10. Wicki, R., Marenholz, I., Mischke, D., Schäfer, B. W., and Heizmann, C. W. (1996) *Cell Calcium* **20,** 459–464.
11. Mailliard, W. S., Haigler, T., and Schlaepfer, D. D. (1996) *J. Biol. Chem.* **271,** 719–725.
12. Seemann, J., Weber, K., and Gerke, V. (1997) *FEBS Lett.* **413,** 185–190.
13. Robinson, N. A., and Eckert, R. L. (1998) *J. Biol. Chem.* **273,** 2721–2728.
14. Lassar, A. B., Buskin, J. N., Lockshon, D., Davis, R. L., Apone, S., Hauschka, S. D., and Weintraub, H. (1989) *Cell* **58,** 823–831.
15. Haglid, K. G., Hamberger, A., Hansson, H. A., Hyden, H., Persson, L., and Ronnback, L. (1974) *Nature (London)* **251,** 532–534.
16. Mandinova, A., Atar, D., Schafer, B. W., Spiess, M., Aebi, U., and Heizmann, C. W. (1998) *J. Cell Sci.* **111,** 2043–2054.
17. Messerli, J. M., van der Voort, H. T., Rungger-Brandle, E., and Perriard, J. C. (1993) *Cytometry* **14,** 725–735.
18. Ilg, E. C., Schäfer, B. W., and Heizmann, C. W. (1996) *Int. J. Cancer* **68,** 325–332.